DATA COACHING: MEASURING THE EFFECTS OF FEEDBACK ON LOW-STAKES TEST MOTIVATION

A Thesis

Submitted to the Faculty

of

Drexel University

by

Nancy Snyder

in partial fulfillment of the

requirements for the degree

of

Doctor of Education

June 2012

Abstract

Data Coaching: Measuring the Effects of Feedback on Low-Stake Test Motivation

Nancy Snyder, Ed.D.

Drexel University, June 2012

Michel Miller, Ph.D.

This study examines the relationships between students' academic motivation, evidence of achievement as measured by assessments and the effects of feedback in mediating effort. Policy makers currently view student achievement is as synonymous with proficiency on standardized tests. Testing students as a means of determining educational productivity is a standard educational practice that has become more significant since the establishment of NCLB. Increasingly, assessments have become accountability measures for schools, a trend that will become even more personal to teachers in Pennsylvania as the state embarks on a new teacher evaluation tool, which uses student achievement data as a "significant factor" in performance ratings. While these assessments then become high-stakes for schools, such measures may carry little meaning for the students themselves. Research has indicated that learning motivation and ability beliefs decline as students reach middle school. Studies have also shown that student test effort impacts achievement scores. Valid interpretation of test scores is dependent on students' giving full effort.

According to expectancy-value theory, examinee effort is a construct of value and expectancies. This mixed-method, quasi-experimental study tested the effectiveness of a feedback protocol centered on coaching conversations with 36 urban middle school students. The intervention sought to increase task importance and therefore student effort. In a pretest-posttest design using the Student Opinion Scale, students rated their examinee effort and the importance they ascribed to a low-stakes test. Results indicated that this feedback had little impact on student reported importance or effort. Comparison with a control group, subject effort showed greater decrease

Analysis of field notes indicated that confounding variables may have influenced these results. Students who reported positive test effort gave several reasons. In addition to ability beliefs that increased their expectancy for success, students named parents and peers as influencing their perceptions of test importance. Most often, however, students attributed positive test effort to their teacher. Cost of time and energy was most often cited as a reason for failing to give full effort, but negative effort was also attributed to teachers. Here students described adversarial relationships and ineffective classroom practices as negatively affecting effort in class and during tests. Confirming studies of classroom goal structure, students reported that classrooms that focused on performance, rather than mastery goals, negatively influenced their academic achievement and ensuing test effort.

The findings of this study are instructive to school personnel who wish to maximize student effort in testing situations. Using the Student Opinion Scale to assess student effort and importance on tests would enhance instructional practices as well as assist in valid interpretation of test scores, particularly in non-consequential test situations. As students most frequently cited cost as the reason for their failure to give full effort, schools should carefully consider what assessments are necessary to give and ensure that curricula are aligned to assessment material. Because much about current testing practices is mandated, schools should carefully consider the frequency and use of non-mandated tests, particularly those that are low-stakes. Finally, qualitative data endorses the use of mastery goals in promoting full student effort through classroom instruction that minimizes energy costs and improves ability beliefs, task value, and expectancies for success.

The Dissertation Committee for Drexel University

certifies that this is the approved version of the following dissertation:

DATA COACHING: INCREASING STUDENT MOTIVATION ON LOW-STAKES TESTS

Committee:

_____

Michel Miller, Supervisor

_____

Joyce Pittman

_____

Mary Jo Grdina

Dedication

This dissertation is dedicated to students who spend their childhoods being measured and to the school personnel who encourage and support them.

Acknowledgements

I would like to acknowledge the students, teachers, and administrators who participated in and supported this dissertation study. I also wish to thank the chair of my committee, Dr. Michel Miller for her advice and support in completing this project. Finally, I am indebted to my husband, David, my daughters Mallory and Haley, and many friends; all who listened and at least feigned interest as I droned on about this topic.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Introduction

As the policy debate continues regarding the impact of the landmark legislation known as No Child Left Behind (NCLB), schools struggle to meet the rigid demands of testing and accountability. The intention of the law, to increase student achievement levels across the country, is commendable, but unrealistic goals have created unintended consequences for teachers and students (Amrein-Beardsley, 2009; Eckes & Swando, 2009). Teachers are faced with ethical dilemmas of whether to narrow the scope of their teaching in favor of improving test scores, or to broaden their instruction to include untested, but critical, skills (Cochran-Smith & Lytle, 2006; Crocco & Costigan, 2007). Some have also questioned the reasonableness universal proficiency expectations since assessment theory describes large-scale achievement patterns as predictably falling on a "bell-shaped curve" and years of testing has reinforced this notion (Davies, 2008).

The threats of accountability measures significantly affect schools. In some cases, teachers make decisions based not on what they feel is best for their students, but what will bring about necessary assessment scores (Cochran-Smith & Lytle, 2006, p. 683). The pressure of accountability as measured by assessments is so strong that cheating and "gaming" practices are often employed to increase test score—sometimes at the expense of actual learning (Amrein-Beardsley, 2009). According to the National Center for Education Statistics, in 2008 more than 32% of teachers reported that they felt concern about their job security because of their students test scores, an increase from 28% in 2000 (United States Department of Education, 2008).

This concern is more pressing in schools that serve high poverty populations. Here, students are more likely to attend schools labeled as "failing"(Murnane, 2007). These are schools that work desperately to meet achievement standards that will keep them from suffering the most punitive measures of school reform: take over or closure. These issues become enmeshed with debate regarding racial equity and access, particularly among minority students. Much discussion has centered on this "achievement gap" (Braun, Chapman, & Vezzu, 2010; Johnson & Kritsonis, 2010), the significant and somewhat stable differences in achievement scores between white and non-white students across the country. NCLB seeks to redress this gap by investing considerable resources and attention to improving teacher quality (Darling-Hammond, 2009), but fails to consider improving student behaviors that impact high-stakes assessments.

Problem Statement

School and student improvement based on rigid, mandated testing is at the foundation of NCLB. Yet after nearly two decades of implementation, testing alone has not been successful in achieving school reform. Far from improving educational opportunities, "low-income students of color have been the primary victims of high-stakes testing policies that determine promotion, placements, graduation, and base school ranking and sanctions on student test scores" (Darling-Hammond, 2010, p. 74). Schools, school districts, and states are mandated to increase school effectiveness; unfortunately some of the unintended consequences of this high-stakes environment have created the opposite.

The reliance on student achievement data to determine school successfulness is predicated on the assumption that student test scores are reliable, yet research has

indicated that as test-taking motivation decreases, so do both test performance and test score validity (Wise & DeMars, 2005). Given the impact that achievement scores have for students and schools, it is important to know that results are an accurate reflection of student knowledge.

Whether a test is considered high or low stakes is largely dependent upon the subject of consequence. A test may be high stakes for one student and low for another. The Scholastic Aptitude Test (SAT) is generally considered high stakes for students as these scores are used by institutions of higher learning to sort them for admission. The same tests are considered low stakes for teachers since they are generally not used in determining specific teacher effectiveness. Similarly, accountability tests typically are of greater consequence for teachers and schools than they are for students. For some students, these tests may carry a personal consequence, but not the formal consequence of a grade. Students are not always aware of the high-stakes nature of these assessments, or how test data is used to determine student education decisions—including course selection, school application processes, and retentions. Without this knowledge, they become disengaged from assessments, resulting in score deflation. Brown and Walberg (1993) noted that high volume of testing causes students to care little about assessments that don't have a direct effect on their grades or other systems of accountability. They postulated that this might also explain poor performance of U.S. students on international assessments as compared to students of other developed countries. Systems of accountability use summative test scores to measure school effectiveness, but students may give minimal effort if they do not value assessment results.

Purpose and Significance of the Study

The purpose of the study was to determine the extent that a coaching intervention would improve student engagement on low-stakes tests. Schools designated as low performing constantly seek solutions that will address the achievement problem, and often rely on remediation classes that are intended to provide students with test preparation strategies or remedial skills as a means of addressing achievement concerns. This approach has the unintended consequence of assigning students to less academically rigorous instruction, potentially contributing to the achievement and engagement problems. Wise (2009) found that despite validity concerns caused by student disengagement from testing, data from these tests is used to determine instructional programming, systems accountability, and accreditation. Wise further contended that threats to assessment validity increased commensurately with the number and percentage of students who do not give their best effort. This variable is hard to determine and has rarely been considered.

Little research is available that describes to what extend student understanding of high stakes assessments and data are related to intrinsic motivation and goal setting. While much research exists on student motivation both in terms of theoretical models and practical application, few studies have considered how schools might use this knowledge to leverage validity of high stakes tests. This study aimed to determine the extent that student motivation on tests could be manipulated by informing students about the uses of assessment data, and by giving them feedback on their progress toward proficiency. The findings may inform school leaders about the effectiveness of coaching as an intervention strategy for improving student engagement on assessments.

Research Questions

This study addressed three central questions:

1. What effect does a data coaching intervention have on student test engagement?
2. What aspects of data coaching do students report as most helpful in increasing task value?
3. In what ways do students report that data coaching affects their attitudes toward low-stakes assessments?

Conceptual Framework

The conceptual framework for this study was based on research of learning motivation, the use of feedback in learning situations, and relatively recent studies regarding examinee effort during low-stakes tests. Of these research areas, learning motivation was the broadest, encompassing both educational and psychological studies. While several theories of motivation are considered in the literature review, this study is best understood through expectancy value theory. Eccles et al. (1983) explained achievement choices as constructs of both the expectancy of success and relative value of the task. Expectancy and value "also influence performance, effort, and persistence" (Wigfield & Eccles, 2000, p. 69). The term "expectancies" describe students' perceptions regarding how well they will perform on tasks and is related to ability beliefs. "Value" is defined by the perceived importance and usefulness of a task, the intrinsic and utility values, and the cost of time and effort (Eccles, et al., 1983). Wise and DeMars (2005) used this theoretical framework to explain the results of their study on examinee effort during low-stakes tests. On such assessments students have little direct consequence for

their performance, diminishing the value they place on the test. The researcher hypothesized that weak value beliefs may explain why some students report low engagement on these types of exams.

This study was based, in part, on the researcher's experience as an educator, literacy coach, and school administrator in an urban school setting. Using data to determine student interventions and to monitor overall school effectiveness has been challenging due to the unreliability of results. While ample assessment points and data management systems provide easy access to a large body of information on groups and subgroups of students, the daily interactions with students often challenges the conclusions that the data suggests. A significant number of students' formative and summative assessments show high and low scores within similarly designed tests and within short time spans.

Informal, anecdotal data collected from students, particularly those whose data show erratic patterns, have revealed that lack of student effort is given as one reason for low scores, particularly when they are accompanied by higher data points. This evidence has brought into question the reliability of the measures. Conversations with students have subsequently revealed that they lack understanding regarding the importance of valid data, how tests are constructed, and how their own data impacts both their current and future school experiences.

The goal of this study was to explore the effectiveness that coaching conversations had on students within a particular setting, and whether student-reported engagement increased on assessments following the intervention The protocols for discussing data with students included informing students about the nature of the

assessments they will be taking, how the data will be used, and how data can be interpreted. Counseling on individual achievement was focused on goal setting, motivation, and accessing supports to increase achievement. Figure 1 shows the conceptual framework used in this study. Coaching feedback was used to mediate student value as a means of increasing examinee effort.

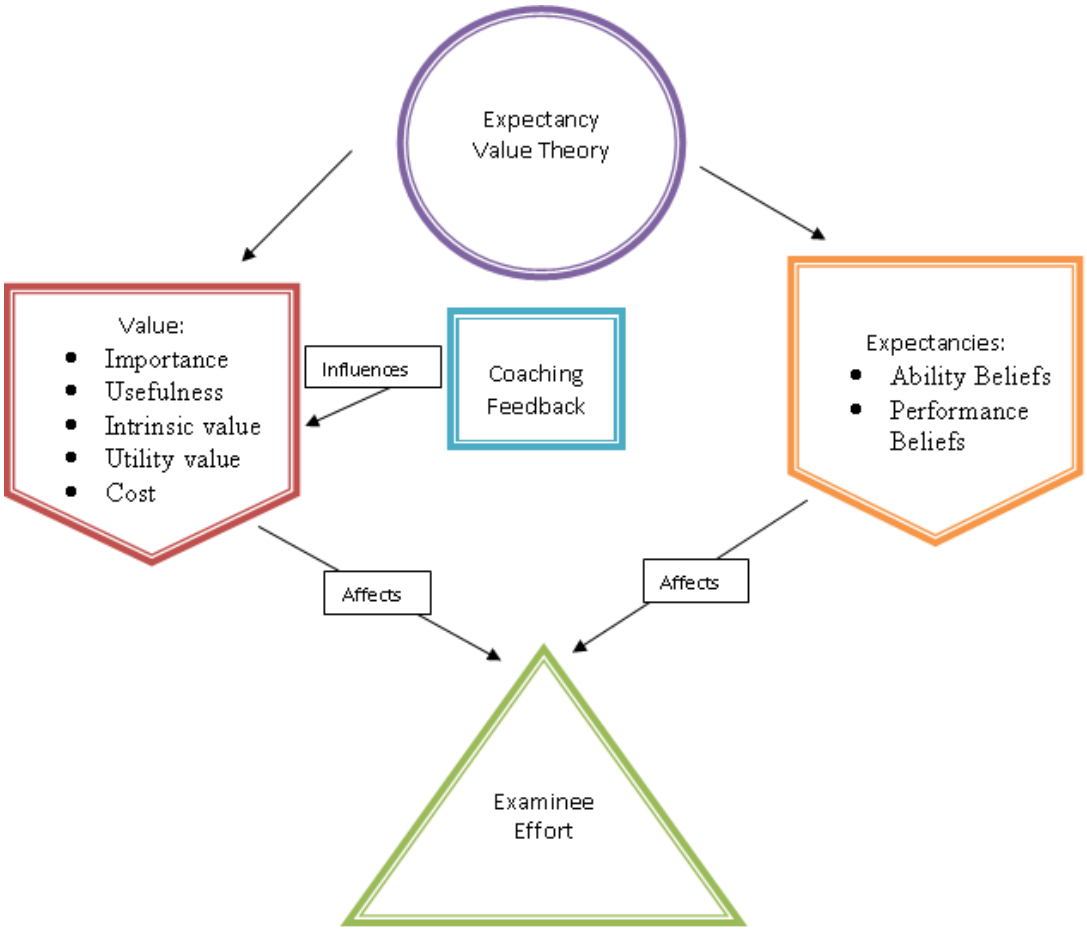


Figure 1. Conceptual Framework

Definition of Terms

Achievement gap: The achievement gap is a term used to describe the disparity in achievement levels, as determined by standardized test scored, between demographic

groups of students. Traditionally, these groups have focused on racial differences; however Ramirez and Carpenter (2005) have found that multiple gaps exist both within and between subgroups, indicating that the gap cannot be correlated with race alone. For the purposes of this study, the achievement gap was defined as the average difference in achievement between students who attend high poverty schools and those who do not.

Data coaching: This is the practice of teaching students about the culture and norms of standardized testing and the accompanying accountability systems. The procedures are both informative and reflective as students review their own data and set goals.

High poverty schools: These schools are identified as such when more than 75% of students qualify for free or reduced lunch programs. A student qualifies for subsidized lunch programs if they come from a household with an income at or below 130% of the poverty threshold for free lunch, or between 130% and 185% of the poverty threshold for reduced-price lunch (United States Department of Agriculture, 2009).

High stakes tests: High stakes tests are those that have serious consequences for students or educators (American Educational Research Association, 2000, p. 24). In public schools many standardized achievement tests are inconsequential to students while educators bear total accountability. Some tests, particularly at the high school level are becoming high-stakes for students. Pennsylvania's Keystone Exams are an example of high stakes assessments for students while the Pennsylvania System of School Assessment (PSSA) would be considered high stakes for educators.

Low stakes tests: In the high-stakes, low-stakes dichotomy, low stakes tests are those that have little consequence to students. In U.S. public schools NCLB has expanded

the emphasis on low-stakes tests as a means of holding schools accountable for the quality of education they provide students (Wise & DeMars, 2005).

Test-taking effort: Brown and Walberg select a more common definition as, "students' propensity to engage in full, serious, and sustained effort on academic tests" (1993, p. 133). Test-taking effort has also been defined by Wise and DeMars (2005) as "a student's engagement and expenditure of energy toward the goal of attaining the highest possible score on the test" (p.2). For the purposes of this study, test motivation was self-reported as "engagement and effort."

Assumptions, Limitations, and Delimitations

While the framework for developmental assets is based on building social and emotional capital in students who have traditionally been described as "at risk", it is assumed that all students can benefit from these strategies. The relationships that students developed with the data coach as a result of this study may have had some influence on student motivation, but were not measured; therefore, they are considered unintended benefits of the study construct.

Due to concerns about test stress, the subjects of this study were limited by age. Data was not collected on students below fifth grade. Modified coaching conversations were conducted, however, with students in fourth grade. Similarly, students beyond eighth grade were also excluded. Results from this study do not reveal whether data knowledge is sustained after a year of implementation and to what extent students will be able to internalize test motivation if they do not have access to their own data in the future.

This study was confined to students who attended high poverty, urban elementary/middle schools. More research is needed to determine if this improvement strategy is equally successful in schools with different demographic characteristics.

Summary

As schools become increasingly accountable for student achievement, strategies for capturing the best scores that students can attain have become imperative. This study determined whether and to what extent informing students about testing and data collection combined with examining student individual data and setting learning goals helped students to become motivated and engaged in what is otherwise low-stakes testing situations. The ensuing review of the literature considered various theories of motivation as they related to both learning and testing environments, the impact of low examinee motivation, and the use of feedback as part of a data coaching intervention.

Chapter 2: Literature Review

Introduction

This study investigated the relationships between students' academic motivation, evidence of achievement as measured by summative assessments, and the effects of feedback in mediating effort. Policy makers currently view student achievement as synonymous with proficiency on standardized tests. While these assessments then become high-stakes measures for schools, Wise and DeMars (2005) found that such measures carry little meaning for the students themselves. The authors claimed that when test-takers were not engaged in the assessment process, both test performance and test validity decreased.

The concern about proficiency is greater at the transition to middle school or junior high. During these critical years, motivation, interest in learning, and effort decline across subject areas. Some studies indicated that these changes are more concerning for African American and Hispanic students (Steinberg, Dornbusch, & Brown, 1992; Warner & Phelps, 2008). While African American students showed similar levels of engagement and enthusiasm for learning in elementary school, by adolescence, they became much less committed to their school experience, seeing "less of a connection between academic performance, their self-worth, and career success than other students"(Warner & Phelps, p. 72). While all schools can benefit from a deeper understanding of student motivation as it relates to achievement, the case is more imperative for schools serving high populations of minority students.

Conceptual Framework

The conceptual framework for this study was based on literature regarding learning motivation, assessment motivation, and the use of feedback with students. For some number of students, test results conflict with expectations. Many students show enough effort that their scores merely confirm what their parents and teachers already know. This is particularly true in younger grades where students are eager to please adults. With some adolescents, however, scores are disappointingly low. With a wide array of data now available through data warehousing, a significant number of students' formative and summative assessments show fluctuation within similarly designed tests and within short time spans leading to questions about the validity of results.

Students have reported lack of effort as one reason for low scores, particularly when these scores are higher at other points. Conversations with students have subsequently revealed a lack of understanding about the importance of valid data, how tests are constructed, and how their own data impacts both their current and future school experiences. This fact is concerning given the importance of test scores as accountability tools. Whereas school staff must focus on achievement and take seriously the tools used to measure progress, students do not all share these concerns. If student motivation impacts test results, what can schools do to ensure that all students give their best effort?

An expectancy-value model of motivation helps explain why students may not give their best effort on assessments. According to this model, expectancies and values are influenced by a student's ability beliefs, their perceptions of the difficulty of the task, their goals, self concept, and affective memory (Eccles, et al., 1983). These expectancies and values then affect effort, achievement, performance, and persistence. Students who

have low competency beliefs may predict that they cannot do well, and therefore will not try. Weak value beliefs would account for students who may have the ability, but either do not see the purpose in doing well or view the cost of time and effort as being too high.

The question is not whether schools can affect student effort, but rather how. Some students, because of low ability beliefs, will require intensive support and encouragement beyond school staff. These students were not the focus of this intervention. This study aimed to address low task beliefs among students who have the ability to perform well but don't, either because they see little use in these assessments, or because they are unwilling to expend the time and concentration. To what extent will coaching students on the purpose and personal value of accountability tests be sufficient in prompting best effort? Figure 2 offers a graphical representation of the ensuing review of literature.

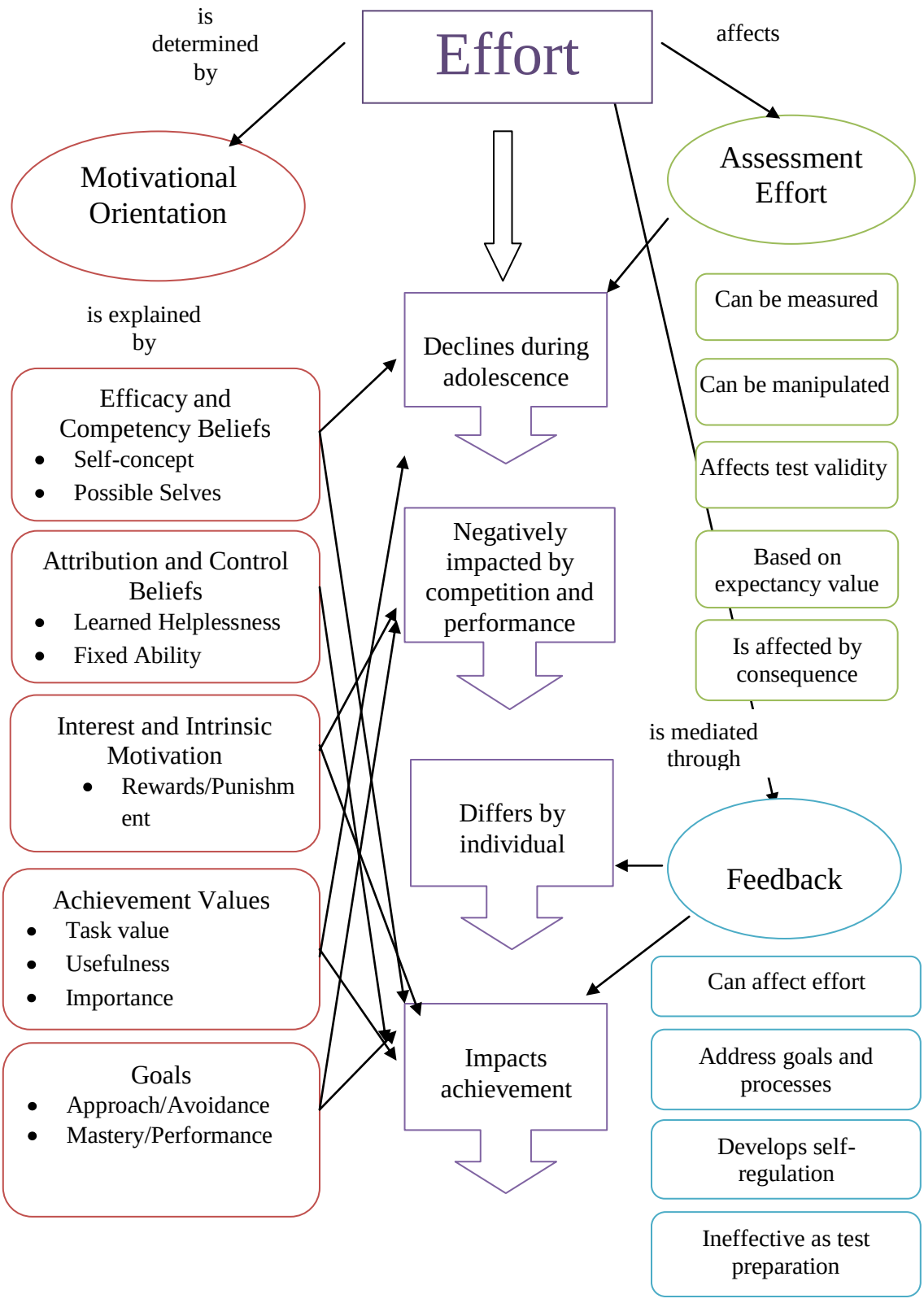


Figure 2. Framework for the Literature Review

Literature Review

In order to better understand how effort translates into demonstrated achievement, this literature review examined current research regarding motivation theories, the relationship between effort and test taking, and the effects of feedback on motivation. Literature on motivation and feedback is instructive in the formation of the data coaching protocol since the intention of this intervention is to improve student motivation and engagement. Literature on assessment motivation highlighted the concern, which is the foundation for the study.

Although student motivation has been studied at all age levels, concern about the decline in motivation in middle school students bears examination. The studies have tested motivational theories with a specific age group and offered insight regarding ways that educators could influence learning motivation during this critical period. Assessment motivation is a subcategory of motivational research that requires separate investigation. This body examined how lack of motivation and test engagement has statistical effects on achievement scores. Findings challenge the assumption that high-stakes accountability testing is valid and reliable. Finally, research studies that examined the effectiveness of feedback are considered as possible ways to address low examinee motivation. This research informed the later development of an intervention protocol that is the subject of this dissertation study.

Adolescent motivation. Social cognitive and organismic models of motivation have attempted to describe the forces behind action including beliefs, values, goals, and interests. Social cognitive models have described processes and activities that explain students' choice, engagement, persistence, help seeking, and school performance

(Pintrich & Schunk, 2002). Organismic models examined how individuals internalize values and regulate their behavior as they interact with their environment (Deci & Ryan, 1996). Between these two models, a wide array of theories regarding motivation and academic achievement exist. The research surrounding these theories is voluminous, and the discussion in this literature review is not intended to be exhaustive. In creating a motivational intervention, the sum of this knowledge is instructive in framing motivation variables and is critical in avoiding actions that might ultimately be antithetical to the desired outcome.

Pintrich (2003) organized the concepts of motivational theories into five constructs: efficacy and competency beliefs, attributions and control beliefs, interest and intrinsic motivation, achievement values, and goals. Efficacy and competency beliefs center on self-concept and a student's ability beliefs (Markus & Nurius, 1986). Attribution theory describes subjects' perceptions of the reason that an event occurred either by ability, effort, or luck. In the context of achievement, failure is attributed to low ability or lack of effort, while success is considered the result of ability or effort (Weiner, 1985). Intrinsic motivation considers whether the impetus for action is located within a person (intrinsic) or is in response to external factors (extrinsic). This motivational orientation in turn has an effect on an individual's self-regulation ability (Deci & Ryan, 1996). While interest describes what children like, students ascribe value to what is useful. Expectancy-value theorists describe motivation as a construct of students' ability beliefs and their perceptions of a task's usefulness, importance, or interest value. (Wigfield & Eccles, 2000). In motivational literature, goals were described as either performance oriented or mastery oriented. In the case of the former, goals are focused on

demonstrating high ability as compared with others and using these comparisons to make judgments of ability and performance. Students derive a sense of accomplishment from doing better than others and exceeding normative expectations. Conversely, mastery orientation describes a focus on self. Students are oriented toward self-improvement and in mastering a difficult task. Satisfaction is derived from interest and in challenge (Meece, Anderman, & Anderman, 2006; Wigfield & Cambria, 2010).

Studies regarding middle school students and early adolescents are critical in explaining the phenomenon of motivational decreases and shifts in orientation at this age. Additional work examining the relationships between gender and ethnicity are important considerations for the sample of this study. Testing motivational theories provides guidance in constructing interventions to engage students in academic achievement striving and assessment motivation.

Competency beliefs and the transition to middle school. Several researchers have studied the effects of changes in student perception and classroom orientation in the transition from elementary to junior high school or middle school. Harter, Whitesell and Kowalski (1992) conducted two studies that contribute to the understanding of this idea. The first considered the effect of transition on students' academic self-concept and motivation. This study surveyed 463 students from across two school districts using several tested tools and found many student perceptions shifted during the transitional year. Half the students maintained similar competency beliefs from the year prior, but half were likely to change their perceptions nearly equally to the positive or negative. Changes in competency beliefs were accompanied by changes in motivational

orientation. Students who felt more competent showed greater intrinsic motivation while those whose beliefs declined were less intrinsically inclined.

The second study in this report considered whether perceived changes in the school environment were related to competency beliefs. The sample for this study consisted of 338 sixth, seventh, and eighth graders in a single middle school. Findings suggest that students who perceive the school as emphasizing performance and competition showed higher extrinsic motivation and higher anxiety than those who did not. Together these studies show that student perceptions of middle school environments as well as competency beliefs have an impact on student motivation.

Competency beliefs and goal orientation. Similarly, Anderman and Midgley (1997) considered the changes in achievement goals, classroom goal structure, and academic competency perceptions of students following a transition from elementary school. A multivariate analysis was conducted on students grouped into high and low ability groups based on a standardized measure of cognitive ability. This analysis considered personal task goal orientation, personal performance goal orientation, perceived classroom performance goal structure, perceived academic competence, and grades.

The findings demonstrate a dramatic decline in student perceived confidence in middle school compared to ability perceptions in elementary school. Classrooms were also reported as more focused on ability rather than mastery and improvement. Interestingly, ability scores remained stable, but grades decreased for most subgroups. This study was one of the first to indicate that learning environments for middle school students were less supportive toward mastery than those in elementary schools: It added

to a significant body of research that tied increased focus on performance with a decrease in motivation.

Concept of possible selves. In their work to link self-concept with motivation, Markus and Nurius (1986) examined student's possible selves, which are conceptualizations of themselves in the past, present, and future. Through quantitative analysis of survey data, this study found that self-image functions as an incentive for future behavior and that while past images have a bearing on possible selves, these images are not static. Building on this work, Anderman, Anderman, and Grieslinger (1999a) examined the relationship between possible selves and goal orientation in early adolescence. They found that images of possible selves were related to achievement, motivation, and grade point average. They also found that performance goal orientation is endorsed for future selves. While this was encouraging given Markus and Nurius's finding that self-concept is malleable, it is also concerning that adolescents perceive that "good students" need to perform well in class rather than master material. This study supported Anderman and Midgley's (1999b) earlier finding that middle school classrooms are more often performance-oriented and strengthens Harter's (1992) claim that middle school environments impact motivation.

Attribution. Attribution theorists consider a subjects' perception of the reason that an event occurred. In the context of achievement, failure is attributed to low ability or lack of effort, while success is considered the result of ability or effort. These dichotomies affect both teacher and student perceptions and may contribute to punishments and rewards that further reinforce high or low motivation (Weiner, 1985).

Learner attribution contributes to "achievement striving", a behavior characterized by working with intensity, selecting challenging activities, initiating learning tasks, and persisting even with the risk of failure. High-needs achievers more readily assign success to personal factors of ability and effort, resulting in feelings of pride and increased likelihood for continued motivation (Weiner, 1972).

Conversely, Maier, Seligman, and Solomon (1968) described the corollary to achievement striving as "learned helplessness." Research conducted with dogs revealed that when subjects receive negative stimulus for a primary learning task, they become conditioned against self-reliance in future tasks. Weiner (1972) applied this term to low achievement learners, since they do not perceive that effort influences outcome. Subsequent studies found that students see ability as a fixed characteristic that they could not control. This view further affects motivation in that ability attribution positively reinforces task success while failure has negative motivational consequences (Weiner, 1985).

Declining value in early adolescence. More recent expectancy-value models have examined task value: the importance, value, usefulness, and perceived cost of an activity. In their study of over 1,800 junior high school students, Wigfield, Eccles, Mac Iver, Rueman, and Midgley (1991) showed that students' value of math, reading, and sports declined as they transitioned to seventh grade. Girls showed a greater value for reading, whereas boys demonstrated higher value for sports. Research based on expectancy-value models also considered how children's ability beliefs affected motivation. In their decade-long longitudinal study, Jacobs, Lanza, Osgood, Eccles, and Wigfield (2002) studied competency beliefs and task values of three subject domains: math, language arts,

and sports. Figure 3 summarizes findings on growth curves for competency belief in males and females in these subject areas across twelve grade spans.

It is notable that confidence declines across years for both genders in all three domains. These findings also indicated gender differences across years and domains. While boys are more confident in their abilities in math in first grade, by twelfth grade no gender discrepancy is apparent. Conversely, boys and girls begin school with similar views in language arts, but this domain becomes stronger for females. Predictably, gender differences are greatest in the area of sports ability. Wigfield and Eccles (2000) asserted three explanations for this decline: (a) as students develop cognitively, they are better able to assess their abilities and become more realistic than they were in early childhood; (b) they become better at interpreting feedback and engage in more social comparisons with others; and (c) school climates often foster a competitive environment between students causing some to lower their achievement beliefs.

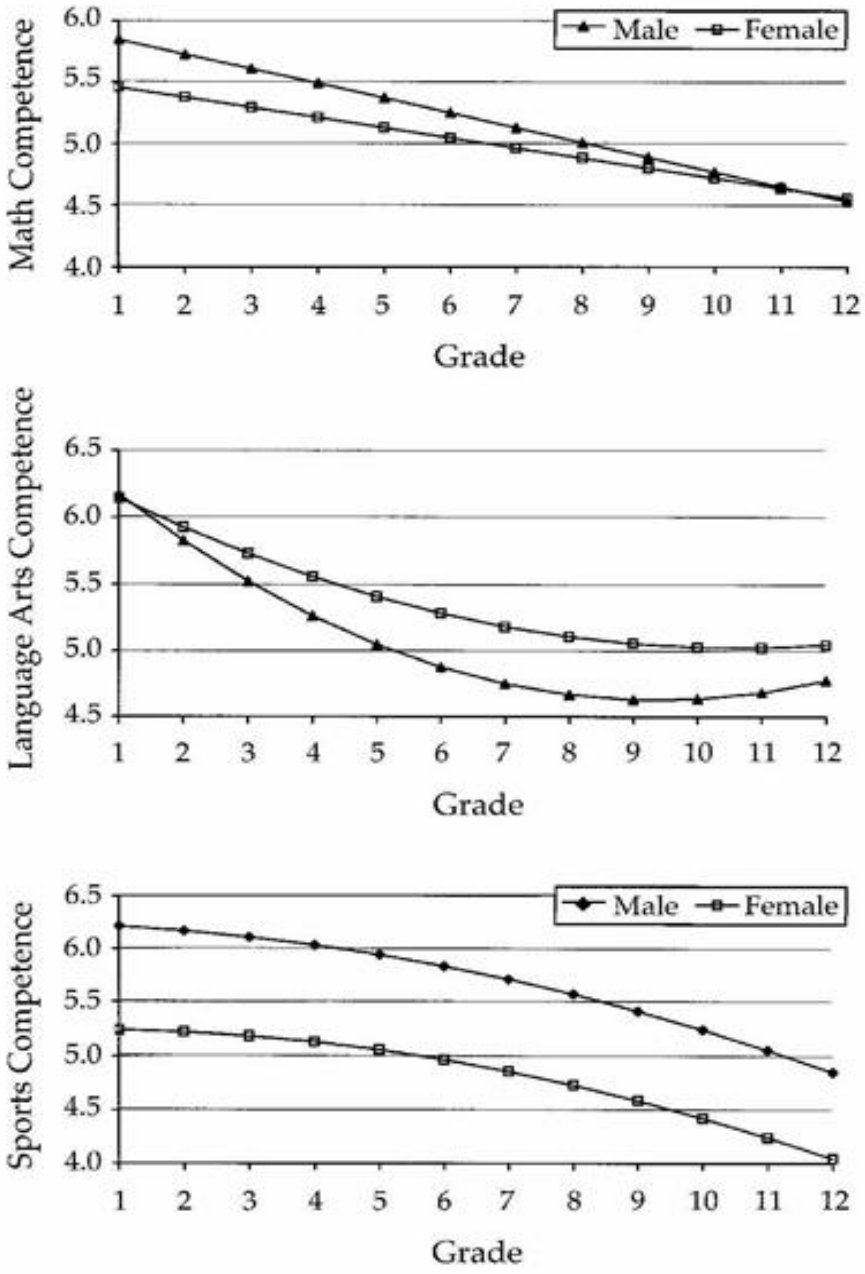


Figure 3. Competency beliefs for males and females in math, language arts and sports from grades one through twelve. From "Changes in children's self-competence and values: Gender and domain differences across grades one through twelve," by J.E. Jacobs, S. Lanza, D.W. Osgood, J.S. Eccles and A. Wigfield, 2002, *Child Development,*73, p.516. Copyright 2010 by Blackwell Publishing. Reprinted with permission

Jacobs et al. (2002) and Wigfield et al. (1991) showed that task value and competency beliefs declined as students transition from elementary to middle school. Ability beliefs and their expectancies for success have been found to be the strongest predictors for academic success; stronger than grades or achievement levels (Wigfield & Eccles, 2000). Efforts to improve student motivation in adolescents must therefore address both student ability beliefs as well as value for academic subjects.

The effect of rewards on motivation. Studies of self determination theory (Deci, Vallerand, Pelletier, & Ryan, 1991) considered human motivation and personality, and showed how motivation and learning is associated to psychological need. Much of the educational application for this theory deemed regulatory process as either one of choice or of external control. In a 1971 study, Deci, (citing White, 1959), noted that motivation could be characterized as innate or learned. This early work considered to what extent external rewards affect intrinsic motivation. In three research trials, Deci tested this question using money as the reward variable in two studies and verbal reinforcement as the variable in the third. His findings suggested that when monetary incentives were given for task completion and then removed, completion motivation decreased. However, when positive feedback and verbal reinforcement was offered as the external variable, motivation increased. This study clarified earlier discrepant findings to assert that when praise and verbal reinforcement are used, the desired behavior persists indicating an internalization of regulation. This finding has become a seminal work in the discussion of intrinsic and extrinsic controls.

Motivational orientation, gender, and ethnicity. In a further study of motivational orientation, Lepper, Iyengar, and Corpus (2005) examined the relationship

between intrinsic and extrinsic motivation and whether differences in orientation could be determined by gender or ethnicity (European American and Asian American). The researchers tested a modification of Harter's (1981) scale of intrinsic versus extrinsic orientation. The questions on the scale were changed from dichotomous ratings to those that would assess intrinsic and extrinsic tendencies on two independent scales.

The findings indicated that rather than being at opposite ends of one continuum, intrinsic and extrinsic motivation orientations can coexist and should be considered orthogonally rather than linearly. Despite this paradigm shift, the researchers confirmed Harter's findings that intrinsic motivation was highest among the youngest children studied (grade 3) and lowest among the older children (grade 8). Ethnic differences were also found in that Caucasian students tended to show a stronger affinity for either intrinsic or extrinsic motivators while Asian students who were intrinsically motivated also showed a strong desire to please the teacher—an external orientation. The authors hypothesized that Asian cultural beliefs regarding authority may not conflict with their otherwise strong internal motivation, but this idea was not explained fully by the results.

Reading motivation, gender, and ethnicity. In their study of motivational orientation and its effect on reading achievement, Unrau and Schlackman (2006) considered the relationship among reading achievement, gender, ethnicity (Hispanic and Asian), grade, and motivational orientation. Analyzing survey results, the researchers found a significant decline in intrinsic motivation across the middle school years for boys and girls. They also found that students with lower reading achievement showed the sharpest decline.

While this study did not reveal significant differences in reading attitudes by culture, differences were noted in motivation and achievement. Intrinsic motivation was positively related to reading achievement for Asian students at a statistically significant level; neither motivational orientation had a direct effect on reading achievement for Hispanics. Unlike Lepper et al. (2005), researchers in this study offered several possible explanations for these ethnic differences. Of interest is Ogbu's cultural-ecological theory (1998), which hypothesized that there are substantial differences between voluntary and involuntary minorities in American society. Fitting with self-determination theory, Ogbu explained that those who are minorities by choice may feel more optimistic about institutions like school and see them as holding promise. Conversely, involuntary minorities view institutions more ambivalently or even with hostility. These views would have a direct impact motivation orientation.

While the results of this research merely confirm what other studies have shown about decreases in student motivation and achievement across the middle school years, Ogbu's cultural-ecological theory and its application in the discussion of these findings is worthy of further consideration. This theory provides a possible explanation for the achievement gap in urban schools across the country. While offering no specific remedy or intervention, the information is critical in understanding how achievement deficits might be explained.

Ethnicity and achievement. During the 1987-1988 school year, Steinberg, Dornbusch, and Brown(1992) studied ethnic differences in school achievement among 15,000 high school students representing wide ranging cultural and socioeconomic conditions. Using an extensive survey that measured several psychosocial constructs and

social influences like peer associations, parenting, and school environment, the researchers hoped to explain achievement differences among White, Asian, Hispanic, and African American youth. Their findings illustrated the complexity of factors that contribute to motivation and subsequent achievement. Acknowledging that a small study was trying to address a large and complex phenomenon, the results bear discussion.

Speaking in general terms, authoritative rather than authoritarian parenting styles and strong peer support for academics were correlated with higher achievement. Caucasian students most frequently reported authoritative parenting and strong peer support. Hispanic students suffered a combination of authoritarian parenting and low peer support while high peer support among Asian students seemed to mitigate their parents' authoritarian style. The benefit that African American students derived from parental warmth was countered by low peer support for achievement.

A second important finding was the motivational effects of fear and optimism about the future. Among all four ethnicities, the researchers found little difference in student perceptions that getting a good education increased future good job prospects. But when they analyzed for fear about the future without a good education, Asian students were far more skeptical in thinking that a good job could result from a poor education. Hispanic and African American students showed the most optimism in this regard. The authors noted that "unwarranted optimism, rather than excessive pessimism, may be limiting African American and Hispanic students' school performance"(p. 726).

Although this study is more than a decade old, it offers important considerations for contemporary students. Whether or not they are attributed to ethnic differences, the findings regarding the perceived link between education and job attainment, the support

of authoritative parenting styles, and positive achievement views by peers are applicable to all students today.

Motivational orientation and educational aspirations. Warner and Phelps (2008) studied the relationship between motivational orientation and educational aspirations of African American middle school students. Researchers conducted a correlation analysis on a relatively small sample (n = 198) using Harter's Motivational Orientation Subscales (Harter, 1981) and a researcher-developed survey of educational aspiration. The results showed a significantly positive correlation between motivation domain and educational aspiration, particularly on the curiosity/interest and independent mastery subscales. No correlation was found between competency belief and educational aspiration. Female students and those high on the independent mastery subscale showed the greatest aspirations for achievement.

In contrast to the abundance of literature documenting a decline in motivation among middle school students and the negative characterization of African American students as "at-risk," this study showed that even students who have relatively low competency beliefs maintain relatively high aspirations. The authors consider whether the relative malleability of student motivation is positive or whether this indicated undue optimism as discussed in Steinberg et al.(1992).

Motivational giftedness. While intellectual giftedness has been an educational construct for decades, a more recent consideration is the construct of motivational giftedness. Students who are motivationally gifted rate high on intrinsic and low on extrinsic motivation scales: They have high competency beliefs, lower anxiety, and greater academic achievement.

Results of a longitudinal study of the development of motivational giftedness indicated that while giftedness may shift in childhood, by early adolescence, students' motivational patterns tend to be stable (Marcoulides, Gottfried, Gottfried, & Oliver, 2008). Using a latent Markov Chain analysis model, researchers analyzed a comprehensive battery of standardized measures of achievement including the Children's Academic Intrinsic Motivation Inventory (CAIMI) collected on 100 children every six months from age 5 to 17. Data analysis revealed that students fell into three categories of motivation: at-risk, intermediate and gifted. Findings indicated that at ages nine and ten, student motivation fluctuates significantly with 57% of nine year olds described as gifted. However, by age 13 motivation declines and becomes more stable.

These findings support earlier work by the same authors that indicated the compounding problem of motivational decline and stability among adolescents. They suggested that interventions to alter motivation should address the concern earlier rather than later.

Synthesis of research on motivation. The decrease in student motivation through the middle school years has been well documented (Harter, et al., 1992; Jacobs, et al., 2002; Marcoulides, et al., 2008; Unrau & Schlackman, 2006; Wigfield & Eccles, 2000; Wigfield & Wentzel, 2007), and studies across several theoretical models reveal the significance of the problem. Early studies that considered self-concept indicated that the way students view themselves both now and in the future can provide motivational intervention. However, attempts to alter self-perception may be better addressed through the field of psychology. In addressing attribution, educators can help students see the connection between effort and ability. Low achievers tend to ascribe learning to ability

(Weiner, 1972), so to the extent that teachers can assist students in developing a sense of efficacy, they can also affect achievement striving.

More instructive may be the findings from studies of goal structure. In addition to personal goal orientation, Anderman and Midgley (1999b) found that classroom cultures have goal orientation as well. Schools or classrooms that value competition and comparisons of ability are more apt to orient students toward performance goals. It is this sense of competition that may explain, in part, why gender differences exist for some domains; particularly in math and sports (Wigfield & Eccles, 2000). Conversely, schools that foster a sense of support, focusing on understanding and effort, are more likely to encourage mastery orientation in students. Mastery goals have been found to have a greater, positive impact on student achievement, particularly among students who lack skill or knowledge (Meece, et al., 2006).

A surprising, but significant concern among African American and Hispanic youth is the finding in Steinberg et al. (1992) that students may be more optimistic about their futures than is warranted based on their achievement. It is curious that students would see a connection between a good education and a good job, but not confirm the corollary that a good job is predicated on a good education. One explanation might be in students' views of their possible selves. Possible selves are "the cognitive manifestation of enduring goals, aspirations, motives, fears, and threats" (Markus & Nurius, 1986, p. 954). These possible selves influence student motivation and goal setting. If students envision futures for themselves that are optimistic, but are not dependent on academic success, then their optimism and motivation may not leverage achievement in school.

Martin and Dowson (2009) showed that students' goals, whether oriented toward approach or avoidance, are influenced by others, including adults and peers. Peer relationships are essential in supporting achievement (Steinberg, et al., 1992), but school personnel can only influence, not control peer support and parenting style. There is, however, a significant association between the relationships that students have with their teachers and the goal orientation they endorse. In this way, positive relationships and a supportive school climate can increase engagement and achievement (Martin & Dowson, 2009). Teachers who value mastery over performance and assist students in the development of self-regulation may be able to help students avoid the decline in motivation and achievement that is pervasive among adolescents today.

In his research on high-low achievement motivation, Weiner (1972) had subjects evaluate their performance on a given task as successful or unsuccessful and attribute the cause of this outcome to luck, effort, ability or task difficulty. He found that individuals high in achievement motivation more frequently ascribed success to high effort and high ability. Meyer (as cited in Weiner, 1972) also found that persons who were identified as low on a motivation scale attributed failure to lack of ability. Meyer further demonstrated that even when high motivators failed, if they attributed the outcome to lack of effort, their future expectancies of success were not impacted. This was in contrast to those who attributed their failure to lack of ability.

Test motivation. A more recent construct in studying motivation is the effect it has on test results. Interest in this field has expanded in past decade, perhaps due to extensive testing required by NCLB. This research applied motivational theories in the context of assessments. Through the lens of task value, assessments can be viewed as

high or low stakes. High stakes tests are those that have personal consequence for students. An example is the Scholastic Aptitude Test (SAT) that determines, in part, admission into college. Even classroom tests can be considered high stakes if they contribute to a student's grade. Task value is highly personal, and students who are intrinsically motivated may give their best effort on assessments regardless of consequence. In contrast, low-stakes assessments are those that have no personal value for students either because scores are not reported, they receive no personal consequence, or because they do not value the measurement or results. Studies in this review considered student effort on low-stakes assessments, its impact on test validity, and whether it could be manipulated.

Increasing effort through test directions. One of the earliest studies of motivation on math scores was conducted by Brown and Walberg (1993). Students (N = 214) from three schools in grades three through eight were given different directions prior to the administration of the Iowa Test of Basic Skills. Experimental and control groups were established at grade level. The control group received standard instructions while the experimental group was read the following script prior to hearing the standardized instructions:

> It is really important that you do as WELL as you can on this test. The test score you receive will let others see just how well I am doing in teaching you math this year. Your scores will be compared to students in other grades here at this school, as well as to those in other schools in Chicago. That is why it is extremely important to do the VERY BEST

> that you can. Do it for YOURSELF, YOUR PARENTS and ME (Brown & Walberg, 1993, p. 134).

An analysis of variance showed significant effect for the experimental group ($F = 10.59$, $p < .01$) and for school, ($F = 3.35$, $p < .05$) but not for grade or gender. The main effect for "school" led the authors to investigate the varying cultures and reach the conclusion that preexisting and pervasive test pressure mitigated the effect of the experimental group at one school.

The results of this study indicated that student motivation has had an effect on test performance, and that motivation can be manipulated. The experimental group showed an increase in achievement scores of .303 SD, representing a 12-percentile point gain. Brown and Walberg showed validity concerns are possible in individual scores and in school comparisons. What is unknown is whether pressure from teachers in the form of test instructions or other means is able to sustain student motivation across years, and what unintended consequences may ensue if students develop text anxiety as a result of this pressure. The design of this study also created uncertainty in the results in that the directions read appealed to both intrinsic and extrinsic motivation (O'Neil, Sugrue, & Baker, 1995).

Increasing effort through monetary rewards. O'Neil, Sugrue, and Baker(1995) compared the effects of monetary awards, task orientation, and goal orientation on a released sample of National Assessment of Educational Progress (NAEP) math scores. Three ethnically diverse experimental groups formed from eighth graders and four groups from twelfth graders ($N = 749$) were given different test directions. Each of the instruction differed in motivational appeal. The monetary group was promised $1 for

each correct answer while the task oriented directions appealed to the students' sense of personal accomplishment and challenge. Terms on the task directions included words like challenging, personal accomplishment, effort, perform, concentrate, and mastering. The goal-oriented directions described the test as a measure of ability and told students that they would be compared with others, and the results distributed to parents and teachers. Words from these directions included ability, compare, and perform. All students were given a companion questionnaire and asked to rate themselves on metacognitive skills to indicate test engagement and effort.

By performing an analysis of variance on the data for eighth grade students, the researchers found a treatment effect for scores ($F = 2.7$, $p = .043$). A post hoc analysis showed that students in the monetary and goal groups scored significantly higher than the task or control groups whose scores were comparable. Additionally, when data from the metacognitive questionnaire was considered, students in the monetary group reported investing more effort than students in any of the experimental groups or the control group. No differences were found among the twelfth grade groups in test scores although the monetary group reported higher metacognitive activity. It is also noteworthy that ethnic differences were not found in metacognitive variables, but were significant in effort and worry. And while mathematical performance was significantly correlated with cognitive strategies and effort, there was a stronger, negative correlation with worry.

The findings of this study support self-regulation theory that effort has an impact on achievement. The relative significance of the intervention in grade eight as compared to grade twelve also supports motivational literature that shows significant decreases in motivation at middle school. The authors report in a pilot study that senior students

suggested a letter of commendation for the highest achieving students would be motivational. This incentive was included in the main study and showed that students who are transitioning to college may be more highly motivated on all assessment measures. Concerning is the high, negative correlation between worry and performance. Test anxiety and the validity concerns demonstrated in performance variance by this study are also cautioned in Brown and Walberg (1993).

Examinee motivation, consequence and item type. Sundre (1999) studied college students to learn about the effects of motivation on two testing conditions: tests that count toward course grades, and test that do not; and also on two test item types: multiple choice and essay. Ninety students were given a test with 30 multiple-choice items and one essay that was described as either "consequential" (graded) or "non-consequential" (non-graded). Upon completion, subjects were immediately given a second test of the same design but of the alternate consequence. A ten-item Likert scale that measured students' effort and task value accompanied each test.

A t-test of motivation and effort ratings for graded and non-graded tests showed significant effect ($F = .79$, $p = .000$) proving that students showed greater motivation and put forth greater effort when they perceived that their scores held consequence. More importantly, a correlation analysis showed that in non-consequential tests situations, motivation accounted for 14% of the variance in test scores. Discrepancies in mean test scores also confirm that students who believed their scores "counted" had higher performance ($F = .62$, $p < .001$). This is especially significant given that the tests were taken by the same examinees. In considering the effect of motivation on essay questions,

findings showed considerable discrepancies between reported motivation during graded and non-graded tests ($F = 1.59$. $p < .001$).

The results of this study quantified several important understandings related to testing and motivation: (1) students are more motivated to perform when the results have a consequence for them, (2) motivation enhances test performance, (3) consequential testing conditions cause higher test performance, and (4) performance on test designs that require more effort, like constructed response and essay questions, is reduced in low-consequence situations.

It is important to note that unlike the other previous studies, this experiment was designed to test examinees under two different conditions eliminating variance according to ability. Despite this construction, significant improvement in test performance was found merely by telling students that their work mattered. By increasing the task value of the consequential test, students reported higher motivation and effort, and these constructs translated into performance. Because this study was conducted on college students, one might conclude that the grade consequence for tuition-paying students might be greater than for K-12 students. However, these results support findings with K-12 populations in both Brown and Walberg (1993) and O'Neil et al.(1995) that motivation affects test performance, and that under specific conditions test effort can be increased.

The effects of incentives on performance in Germany. In a study that examined the effects of student effort on low-stakes assessment performance, Baumert and Demmrich (2001) tested the effects of three motivation variables: (1) promise of feedback, (2) consequence to grade, and (3) payment for higher than mean scores on

student test motivation. The Programme for International Student Assessment (PISA) is a low-stakes test that is voluntary and anonymous. The results, however, have policy and political ramifications much like those associated with NCLB. Therefore, best effort is important to securing valid results.

Ninth grade students (N = 467) were chosen from a single city. In Germany, students are assigned to one of three tracked programs based on achievement after grade 6 (p. 459). Students in this sample represented the highest and lowest tracks. Motivational appeals differed and were presented through the reading of directions as in the Brown and Walberg (1993) study. In the feedback group, students were promised that their teachers would get information about their performance to share with them. As was the design in other studies (O'Neil, et al., 1995; Sundre, 1999; Wise & DeMars, 2005) a motivational survey using a Likert scale measured student effort, task value, and goal orientation among other variables and was collected immediately following the test session.

In an analysis of both motivational survey and student achievement scores, results showed no significant effect for type of motivational appeal. Although higher track students performed better than lower track students, this was expected. More importantly, all students showed statistically similar performance when compared to within track peers. In terms of utility value, girls and higher track students showed significantly higher value for the test. Monetary motivation showed higher utility value for low-track students, while variables that appealed to intrinsic factors was stronger for high-track students. Goal orientation also differed by school in that high-track students indicated greater mastery desire, although this did not have a motivational effect in the feedback group.

In a discussion of the results, the researchers stated that little new learning emerged from the study, merely confirming the findings of other studies. Test directions were not able to have a measureable effect on student performance or effort. The study did reveal some interesting differences among higher and lower track students—information made possible by Germany's unique high school organization.

A meta-analysis of motivational effects on test performance. In their meta-analysis of studies that examined motivational effects on test performance, Wise and DeMars (2005) examined the effect size of 12 studies. Each was designed with a "motivated" or experimental group and a "less motivated" or control group. These labels were constructed for ease in comparing studies and were not intended to indicate that no one in the control group was motivated. Rather, the motivated group was labeled as such to describe the intervention.

Early studies dating to the 1940s and 1950s were conducted on workers, not students, and are less significant than the other ten studies conducted on students. It is interesting to note that among the educationally related tests, one was conducted in 1981; the remaining studies were all published between 1992 and 1999. Effect sizes for the ten studies ranged from 0.07 to 1.49. The largest effect size was derived from an experiment to pay second graders for performance on standardized achievement tests while the lowest effect size tested the monetary effect on high school seniors (O'Neil, et al., 1995). Increased motivation was tested through appealing to student performance goal orientation, increasing consequence, and offering extrinsic rewards. More than half the studies considered the effects of motivation on math scores.

In addition to examining effect size, the authors also considered whether motivational surveys, the most commonly used tool for correlating motivation with performance, could provide valid scores. They concluded that although students were self-reporting on motivational scales, the varied results could be correlated to the experimental manipulations and therefore were "valid indicators of student test taking-effort"(p. 8).

This study is critical when examining the small body of literature related to assessment motivation and student effort. The authors showed that motivated examinees demonstrated higher performance than "non-motivated" control groups by an average of more than 0.5 SD. Equally important is the discussion regarding the use of student self-reporting measures and the conclusion that they can produce valid data.

Response time based measures. Building on the findings from his meta-analysis, Wise (2006) endeavored to show examinee lack of effort through data collected on computer-based assessments. Although he endorsed self-reporting as a valid means of assessing motivation and effort (Wise & DeMars, 2005), this study was designed to more objectively measure the construct through the use of response time.

Assessments that are administered by computer offer researchers unique opportunities to measure student responses. By analyzing the amount of time students spend on each question and comparing it to the time all examinees take, patterns of solution behavior and guessing behavior emerge. Using these data, an index of student effort termed RTE (*response time effort),* can be quantified and questions that elicit high solution behavior can be identified. These items are said to have RTF, *response time fidelity* (Wise & Xiaojing, 2005).

Applying these indices, Wise tested the correlation between RTF and test item characteristics and examined the effects of guessing on test statistics. In terms of items, findings indicated that item length (longer items) and placement (later items) elicited the highest guessing behavior particularly among those that demanded extensive reading. In a second analysis, Wise removed the test items where students had guessed and rescored the means for each question. This filter affected 26% of the sample and therefore had a negative effect on internal reliability from a score of .88 before filtering to a score of .75 after. Correlating filtered scores with SAT verbal and math subtest scores and university GPA demonstrated increased validity for the filtered scores from .36 to .39 for verbal, from .12 to .15 in math, and from .24 to .25 in grade point average.

Wise demonstrated that student effort could be measured objectively through assessing response time. While this was an important finding, it is only applicable to assessments that are computer-based and are designed to gather such data—a significantly limiting factor. More widely applicable was his statistical analysis of the effects of low effort on test validity. In that study and in subsequent literature, Wise (2009) described the use of "motivation filtering" in a wide range of testing situations and its use in improving test validity.

Test motivation and intelligence testing. A recent meta-analysis of 46 randomized experiments compared intelligence quotient (IQ) scores identified under incentivized versus standard testing conditions (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). The analysis showed that incentives increased IQ scores by a mean of 0.64 SD but this effect was moderated by IQ score, having a greater effect on students with below-average baseline IQ (0.96) than those with above-average IQ (0.26).

This indicated that individuals with higher IQs demonstrated greater test motivation and less variability in scores.

In addition to indicating that student motivation is a factor in determining IQ, the discussion in this study reiterated the validity concerns associated with low examinee motivation. The analysis confirmed that motivation could be manipulated, and also revealed a new association between lower IQ and lower motivation. Low motivation has been examined in the context of achievement and low-consequence testing where standardized directions often limit encouragement during testing. But when administering IQ tests, psychologists **are** permitted, even expected, to encourage best effort. It is surprising to learn that even one-on-one testing situations with encouragement can fail to illicit best responses.

Synthesis of research on test motivation. The culture of school accountability has dramatically increased the number and types of tests that students are required to take. Sundre (1999) considered the effect of low-stakes tests—ones that are not included in grading or hold little or no consequence for examinees—on motivation and test performance. Sundre determined that student effort and performance increased when the test became high-stakes. Several studies also suggested that results of low stakes tests may be underestimating student knowledge (Baumert & Demmrich, 2001; Duckworth, et al., 2011; O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005; Wise & DeMars, 2005). O'Neil framed the problem as a education policy concern stating that test scores should be interpreted not as evidence of student knowledge, but of "what they know and are motivated to show us"(2005, p. 154).

This statement illustrates the concern regarding validity of assessments where student effort is low. Motivational filtering, removing test scores from unmotivated examinees, has been shown to improve test validity. In cases where high numbers of test scores were removed due to excessive lack of effort, reliability decreased. More often the reliability coefficient estimates showed little change (Wise, 2006; Wise & DeMars, 2005).

Motivation has shown to be correlated with test effort and is better understood through an expectancy-value model where students who have low ability beliefs or low utility value for such assessments might consider the cost in time and effort is too high (Wigfield & Eccles, 2000; Wise & DeMars, 2005). Several studies found significant, positive correlations between test effort and test performance (Brown & Walberg, 1993; Duckworth, et al., 2011; O'Neil, et al., 2005; Sundre, 1999; Wise & DeMars, 2005). Motivated students outperform their counterparts by more than one-half standard deviation. Wise and DeMars also indicated that motivation was not correlated to ability in that higher ability students are as likely to report low motivation as lower ability students (2005, p. 7). It is encouraging that many studies indicated that modifying test formats and offering incentives can improve motivation, although the practicality and sustainability of this practice is unknown.

Assessment feedback. In addition to improving motivation and effort through incentives and changes in test formats, Wise (2009) suggested the use of feedback and cites an unpublished dissertation by V. Wise (2004) which found that student effort was not impacted by the promise of feedback. Wise named this strategy as a future research need.

Searches for studies using the keyword "feedback" revealed a large body of research related to the use of feedback in a variety of instructional and work contexts. Limiting the search to "feedback and assessment" narrowed results further; however, the majority of studies were related to providing students with the results of classroom-based assessments and the place of feedback in an instructional loop. The following review includes studies that examine the use of effective feedback in instruction and research about the effects of test preparation activities that are broadly termed as "coaching."

The power of feedback. Hattie and Timperley (2007) examined the effectiveness of feedback by comparing the general effect of schooling derived from an earlier synthesis of over 500 meta-analyses to a large meta-analysis of feedback effects. The average effect size of schooling (F = 0.40) was nearly half the average effect size of feedback (F = 0.79), earning it the designation as one of the greatest influences on achievement. Overall effects were further developed in an examination of feedback types ranged from simple teacher utterances to goal setting. The analysis also considered the feedback frequency and context.

From these data the authors suggest a model for effective feedback that (1) sets goals, (2) measures progress to those goals, (3) determines a plan of action (p. 86). The authors further emphasized the connection between this model and research on the importance of self-efficacy and self-regulation to achievement motivation, noting that feedback can be arranged into four types: (a) task or product feedback, (b) process feedback, (c) self-regulation feedback, and (d) personal feedback. Of these, personal feedback is the least helpful for academic goals, while process and self-regulating feedback have the greatest potential for improving students learning.

In a discussion specific to the use of feedback with assessments, the authors stated the importance of feedback not only to redirect learner behaviors, but also to provide teachers with information about their teaching. One limitation in this regard is the assumption that references to assessments indicated classroom-based tests and not the summative, low-stakes assessments that are the focus of this dissertation study. However, there is reason to belief that the positive effects of feedback that are noted in classroom situations could be applied to summative assessments as part of a benchmarked assessment cycle.

Instructional effects of feedback. In a meta-analysis of 40 studies based on an experimental design, Bangert-Drowns, Kulik, Kulik and Morgan (1997) examined the effects of five feedback treatments in a variety of instruction settings. Forty studies provided 58 effect sizes with of mean ES of 4.08. Interestingly, some studies showed negative effect sizes, but a descriptive analysis to explain these results yielded new understandings about the conditions of effective feedback. All studies in which the feedback types were found to support student "mindfulness" showed positive results. The researchers distinguished this type of feedback as different from simply cuing the learner about the correctness of response and use of punishment or praise. While this study examined the effectiveness of feedback through experimental models, the discussion focused on the effect of feedback during instruction. While there was some reference to the connection between instruction and how it prepares students for testing situations, it sheds little light on how feedback can be used to increase motivation on low-stakes assessments.

The effects of coaching student for achievement tests. In an earlier study, Bangert-Drowns, Kulik and Kulik (1983) conducted a meta-analysis of 30 studies to review the effectiveness of coaching programs on achievement tests. The term "coaching program" described a wide range of activities implemented prior to standardized tests that prepared students for the experience and did not refer to a product.

The central questions considered conditions for effectiveness, program features, and whether aptitude or achievement tests are more likely to be influenced by coaching programs. The studies selected were based on coaching models designed to affect scores on tests that assessed specific content, broad achievement tests, or tests of intellectual skill. These activities fell into categories of (a) test-taking orientation and practice, (b) more extensive programs that included drill and cramming, and (c) general instruction on cognitive skills.

The results of this meta-analysis showed that in 25 of the 30 studies, coaching had a positive effect on scores. Although the mean effect was relatively small (raising scores by 0.25 SD) some programs showed considerably higher effect sizes. For example, programs that were based on increasing cognitive skill showed the greatest promise, ES = 0.66, while test-taking orientations that were of short duration were also effective, ES = 0.17.

Many of the studies here were specifically designed to improve students' domain ability or to provide them with test taking strategies. Student motivation on the tests was not considered or measured. One interesting finding was that effectiveness was correlated to the duration of the coaching intervention. This reinforced the findings in other studies (Bangert-Drowns & Kulik, 1991; Hattie & Timperley, 2007) that suggested feedback

(considered here as a type of coaching activity) is most effective when it reaches student self-regulation: a process that involves a greater time commitment.

Coaching for low-stakes testing in Germany. Due to the lack of studies that examined the effect of feedback on low-stakes testing, the search was extended to include applicable studies from other countries. Based on Bangert-Drowns, et al. (1983), Brunner, Artelt, Krauss, and Baumert (2007) conducted a quasi-experiment that examined the results of pretest and posttest achievement scores to determine the effects of teacher-designed coaching, defined again as test preparation activities, on low-stakes achievement results in reading and math. The research also considered the effect of pretesting on achievement results. Pretesting effects were not found to be significant and even when combined with three hours of coaching intervention, very little effect was discovered. Only high-achieving students had modest increases in reading scores.

This study closely aligned with the central questions examined in this dissertation study. The sample populations are similar in age and represent both high and low achieving groups. The dependent measure is similar to the PSSA in terms of purpose and consequence, both to students and teachers. Because this study was conducted in Germany, however, the cultural differences both in students and in academic settings cannot be known. Additionally, the dependent variables of coaching and pretesting are significantly different from the concept of feedback. Despite these differences, the authors cited concern about the effects of student effort on low-stakes tests, thus these findings become important given the lack of applicable research.

Synthesis of research on feedback. Research in the area of feedback as it applies to low-stakes testing is critically limited. Studies in this area most frequently address the

use of a feedback cycle in instructional settings; none were found that examined how this tool could be used to improve assessment effort. Even forms of assessment preparation like testing strategies and orientation were found to have little impact on demonstrated achievement. It is interesting that both German studies (Baumert & Demmrich, 2001; Brunner, et al., 2007) expressed the same concern about student effort and its impact on low-stakes testing. The later study even referenced NCLB as an example of the political and policy concerns that Germany shares with the United States. Despite wide searches for studies that employ feedback to increase student goal orientation and test utility value, little is known about the potential for this type of intervention.

Synthesis of Literature

This review examined scholarly studies related to three areas of this dissertation study: motivation theory, assessment motivation, and feedback. Several studies and meta-analyses show that student effort can threaten test validity (Brown & Walberg, 1993; Duckworth, et al., 2011; O'Neil, et al., 2005; Sundre, 1999; Wise & DeMars, 2005). In fact, Wise (2009; 2005) has shown that by removing test scores from unmotivated students, test validity actually increases. Another important finding is that student motivation can be measured and quantified in reliable ways (Wise & DeMars, 2005; Wise & Xiaojing, 2005). Of these methods, self-reporting surveys have most frequently been used to measure a variety of student motivational orientations (Baumert & Demmrich, 2001; O'Neil, et al., 2005; Sundre, 1999).

Of critical importance is the examination of how student effort can be manipulated and even improved. Studies indicated minimal results in manipulating test instructions to appeal to various student motivators (Baumert & Demmrich, 2001; Brown

& Walberg, 1993). O'Neil et al.(2005) showed a small effect in rewarding grade 8 students with $1 for each correct answer, but this incentive had little measurable impact on grade 12 students. Baumert and Demmrich (2001) also found that payment for scoring above the mean was not effective in producing higher achievement at grade 9. However, Sundre (1999) showed that increasing the consequence to students was significant both in terms of motivation and achievement.

Motivational studies help explain these phenomena. According to the expectancy-value theory, student effort is determined both by students' ability perceptions and by the value they place on tasks (Wigfield & Eccles, 2000). Self-determination theory also explains the impact of rewards on effort (Deci, 1971; Harter, 1981; Lepper, et al., 2005) and shows the connection between internalizing rewards and self-regulating behavior (Deci & Ryan, 1996). Several studies showed that both ability perceptions and task value decline as students leave elementary school (Anderman, et al., 1999b; Harter, et al., 1992; Jacobs, et al., 2002; Lepper, et al., 2005; Unrau & Schlackman, 2006; Wigfield & Eccles, 2000; Wigfield, et al., 1991). Marcoulides, Gottfried, Gottfried, and Oliver (2008) elevated the concern by finding that not only does effort decline at adolescence, it also becomes more stable. One possible reason for this decline is that middle schools may promote competition and appeal to performance goals rather than mastery goals, (Steinberg, et al., 1992; Wigfield, et al., 1991) and that important peer influences at this age do not always support achievement (Steinberg, et al., 1992).

Motivation was also found to differ by gender and ethnicity. Girls tend to find greater task value in assessments (Brunner, et al., 2007) and greater competency beliefs in language arts (Jacobs, et al., 2002). Of particular interest, however, is the similar rates

of decline in student competency beliefs in reading and math as students age (Jacobs, et al., 2002). Ethnic differences in motivation were found to be a function of parenting styles and peer support for achievement (Lepper, et al., 2005). Additionally, Steinberg et al. (1992) discussed the concept of "unwarranted optimism" among African American students who view their future lives positively regardless of current achievement.

Although solutions to the problem of middle school motivation are available (Anderman, et al., 1999b), less is known about how to increase effort on low-stakes assessments. Because feedback is used effectively in a variety of instructional settings with good results (Bangert-Drowns & Kulik, 1991; Hattie & Timperley, 2007), this tool was examined as a possible means to address test effort. Two studies were reviewed that examined the effects of coaching on student test performance (Bangert-Drowns, et al., 1983; Brunner, et al., 2007). Both found few positive effects of coaching, although coaching is defined her to describe test preparation, not feedback. Baumert and Demmrich (2001) examined that the promise of test results feedback to students' schools is not significantly motivational for German grade 9 students, but the design of this study measured this variable through the manipulation of test instructions, a manipulation that had earlier been found to be unreliable in increasing test effort (Brown & Walberg, 1993; O'Neil, et al., 2005).

Numerous studies in the past decade have demonstrated that student test effort is a threat to test validity. The majority of this work has examined this phenomenon in college testing (Sundre, 1999; Wise & Xiaojing, 2005). Four studies examined test motivation among K-12 students (Baumert & Demmrich, 2001; Brown & Walberg, 1993; Brunner, et al., 2007; O'Neil, et al., 2005), two in Germany and two from the United

States. One of these studies considered the effects of coaching on achievement testing (Brunner, et al., 2007) while the other three manipulated test directions with varying motivational appeal to test the impact on student effort.

While these studies discuss the relative difficulty of affecting effort and achievement, it is important to note the lack of personal investment in the students they study.

During this review, no research was found that examined the impact of providing individualized assessment feedback on effort. An intervention coaching protocol is distinguished from other coaching programs as one that gives students regular, personal feedback about their performance on benchmark assessments and also requires them to set goals and plan strategies to achieve those goals—processes that are named as essential to effective feedback (Hattie & Timperley, 2007). This intervention endeavored to increase effort by addressing student competency beliefs, task value, and self-regulation through a protocol of data education, regular feedback, and goal setting. More research is needed to understand whether coaching students on the purpose and personal value of accountability tests can prompt student effort.

This study addressed this gap through inquiry of these central questions:

1. What effect does a data coaching intervention have on student test engagement?
2. What aspects of data coaching do students report as most helpful in increasing task value?
3. In what ways do students report that data coaching affects their attitudes toward low-stakes assessments?

Chapter 3: Action-Oriented Research Methodology

Introduction

Landmark legislation, NCLB, required that school effectiveness be determined primarily through the use of assessments. School institutions, therefore, have become increasingly concerned with student performance on these measures. While the punitive nature of the law makes the assessment high-stakes for schools, students are rarely subjected to sanctions for their performance, creating a low-stakes testing situation that can negatively influence student motivation.

Student disengagement from any assessment can affect both the reliability of the score and the validity of the assessment itself (Sundre & Moore, 2002). Test takers are not able to score above their ability without "cheating" and security measures on high-stakes tests mitigate this possibility. What is unknown is to what extent students underperform. One indication that this phenomenon may be occurring is a fluctuation of scores across a year. Schools that use benchmark testing, the practice of assessing students periodically against an end of year standard, are able to identify these students if data indicates fluctuation of concept mastery, especially when questions are similarly worded. In order to both increase the validity of the assessment and to fairly measure progress for school systems of accountability, leaders must find ways to ensure that student motivation and engagement are optimized, especially during accountability testing.

This study was designed to discover to what extent student intrinsic motivation on low-stakes testing is increased through the use of feedback, goal setting, and reflection. Low-stakes, rather than high-stakes tests were used for two reasons. First, since

consequences are inherently lower on low-stakes tests, improving effort is more difficult than on high-stakes measures. Second, testing protocols are highly prescribed in high-stakes testing situations, prohibiting any alteration. Using the Student Opinion Scale (Sundre, 2007), students in experimental and control groups assessed their effort following low-stakes benchmark tests. These data were used to determine the effectiveness of a protocol designed as an intervention for student test effort. Survey data from intervention participants were used to determine what aspects of coaching students found most motivating and to assist in evaluating effectiveness and implementation. Qualitative data in the form of field notes were coded and analyzed to support the findings. These central questions formed the basis for research:

1. What effect does a data coaching intervention have on student test engagement?
2. What aspects of data coaching do students report as most helpful in increasing task value?
3. In what ways do students report that data coaching affects their attitudes toward low-stakes assessments?

Site and Population

Population description. This study aimed to determine the extent that data coaching was effective in improving test motivation for students. Poverty is correlated with decreases in school achievement (United States Department of Education, 2010). Schools that serve high poverty populations are significantly impacted by the achievement imperatives of the law. High poverty schools are those where more than 75% of students qualify for free or reduced lunch programs. In order to qualify, a student

must be from a household with an income at or below 130% of the poverty threshold for free lunch, or between 130% and 185% of the poverty threshold for reduced-price lunch. Table 1 shows that nearly 20% of elementary schools in the United States serve student populations that are considered high poverty, and that Black and Hispanic students are likely to attend schools where more than half the students are living in poverty. In addition to the effects of poverty, the literature also indicated that achievement declined as students approached middle school (Jacobs, et al., 2002; Wigfield, et al., 1991).

A convenience sample of students in seventh and eighth grade who attend a high poverty school in the state of Pennsylvania served as the target population for this study. Because of the sensitive nature of discussing achievement data with students, it was imperative that school districts trust the intentions of the researcher. Since the researcher worked within the school system, she has preexisting relationships: necessary when working with students. Additionally, parents and school officials are more at ease in giving access to their students. The researcher met with students and contacted parents directly to gain permission for students to participate in the study. Procedures for gaining parent permission are detailed in the procedure section of this chapter.

Disengagement in tests, particularly for African American students, begins in early adolescence, and by eighth grade becomes predictive of future school success (Warner & Phelps, 2008). Students in these grades who take the Pennsylvania System of School Assessment (PSSA), regardless of gender or ethnic group, were eligible to be included in this study. Students who had been in the country for less than a year were excluded from the study, since they were excluded from the reading assessments. Language barriers could also have impacted fidelity to the intervention. For similar

reasons, students who had limited communication capabilities as described in their IEP (deaf and hard of hearing students), or those whose IEPs qualified them for alternative PSSA assessments were not included. Classes averaging twenty produced a population of roughly eighty students. A similar population of students from a neighboring school served as a control group.

Table 1. *Percentage of Students in Schools Eligible for Free or Reduced-Price Lunch*

Race/ethnicity	Number of students	Total	0–25	26–50	51–75	76–100	Missing
			Elementary				
Total	**31,176,444**	**100.0**	**25.4**	**25.6**	**23.5**	**19.9**	**5.5**
White	16,713,023	100.0	35.5	32.2	20.3	5.1	6.9
Black	5,270,943	100.0	8.3	16.9	28.5	40.0	6.2
Hispanic	6,950,840	100.0	12.1	16.8	27.9	41.5	1.6
Asian/Pacific Islander	1,494,329	100.0	39.1	24.0	18.8	15.2	2.9
American Indian/Alaska Native	359,663	100.0	12.2	23.9	31.6	28.1	4.3
			Secondary				
Total	**16,112,947**	**100.0**	**36.2**	**33.8**	**17.3**	**6.5**	**6.3**
White	9,386,497	100.0	47.1	35.4	8.9	1.2	7.4
Black	2,632,525	100.0	15.0	31.6	31.1	15.0	7.3
Hispanic	2,989,287	100.0	19.0	31.9	31.4	15.4	2.3
Asian/Pacific Islander	801,687	100.0	44.9	30.7	15.9	5.3	3.1
American Indian/Alaska Native	193,173	100.0	23.5	33.0	24.7	14.5	4.3

Note: This table (United States Department of Education, 2010) shows that high poverty schools are primarily attended by Black and Hispanic subgroups, particularly at the elementary level

Site description. School A was identified as the intervention school, and was one of five schools in this school district that served middle school students. School A was a school-wide Title 1 School serving 439 students in grades PreK-8. The student body was

64% are African American, 20% Hispanic, 11% Asian, and 4% Caucasian. Twelve percent of students received English Language Learner (ELL) services while 22% of students required special education services. During the 2010-2011 school year, 81% of its students qualified as low-income; the overall school district poverty rate was 82%.

School A was comprised of 20 homerooms, two at each grade level, preschool through eighth grade, two part-time emotional support classrooms, three additional special education teachers, and two ELL teachers. One reading specialist provided interventions, and there were no classroom aides or paraprofessionals. The preschool classrooms had recently been transitioned to the Head Start Program, having previously been operated by the school district for ten years as a reform effort. The office staff included a secretary and one office assistant who managed student data systems.

The school was considered to be in Corrective Action II year one for the 2010-2011 school year. This designation under NCLB indicated that the school had not made adequate yearly progress (AYP) in all subgroups for two consecutive years in more than six years. However, the school had shown continued growth in nearly all disaggregated subgroups between the years 2007-2010. In 2008-2009, School A made AYP through Safe Harbor in all subgroups. The following year it failed to make AYP due to lack of sufficient progress with special education students in reading. The school was successful in the other twenty-three categories, and students met the state proficiency goal in math.

While the number of English language learners (ELL) students had remained somewhat constant over the past few years, the percentage of students who were level one and level two according to ACCESS (Assessing Comprehension and Communication in English State-to-State for English Language Learners) had risen dramatically.

ACCESS is an English language proficiency assessment given to students in all grades who are identified as English language learners. It is given annually in World-Class Instructional Design and Assessment (WIDA) Consortium member states to monitor students' progress in acquiring academic English. These are students who demonstrate beginner levels of English proficiency and therefore demand a much higher level of support, particularly in grades four through eight. Achievement results for ELL students are counted toward the school's AYP status after they have been in the country for one year.

School B was identified as the control school and had the same grade configuration as School A. School B was also a school-wide Title 1 School serving about 600 students in grades PreK-8. During the 2010-2011 school year, 87% of its students qualified as low-income; the overall school district poverty rate was 82%. School demographic data for School B showed 77% African American, 16% Hispanic, 1% Asian, and 4% Caucasian. One percent of students qualified for ELL services while 26% received special education services.

School B had 28 homerooms, three each for grades first through eighth and four preschool classes. Support staff included three part-time learning support classrooms, three additional special education teachers, and one ELL teacher. Two reading specialists provided interventions, but there were no classroom aides or paraprofessionals except in the part-time special education classes. As with School A, the preschool program had recently been transitioned to the Head Start Program. The office staff included a secretary and one office assistant who managed student data systems. English language learners

(ELL students) had recently returned to School B. Prior to this school year, these students had been assigned to cluster schools like School A to receive this service.

The school was considered to be in its first year of School Improvement I for the 2010-2011 school year. This designation under NCLB indicated the school had made adequate yearly progress (AYP) in all subgroups for two consecutive years. However School B did not earn this distinction in 2009-2010. In 2010-2011, School B made AYP in two of ten subgroups.

Site access. The researcher had access to both sites through an agreement with the principals and with the superintendent. A formal letter detailing the research study was sent to the superintendent and principals for signatures of consent to conduct research within the schools. A letter was provided to the parents of all students in seventh and eighth grade describing the purpose and timeline for the study, ways that their children may be involved, and a request for permission. In both schools, voluntary teacher contacts were recruited to assist with test administration and with collecting permission slips.

Research Design and Rationale

This study was designed as a mixed-method, quasi-experimental study using pretest and posttest scores to determine the extent that the use of assessment feedback had on improving student engagement on low-stakes assessments. The experimental nature of this design required two samples of students: one who received a coaching intervention (School A), and a second that served as a control (School B).

According to Creswell (2008), experimental designs are the "most rigorous and strong experimental designs," but quasi-experimental designs frequently occur in

educational settings due to unavailability of randomized grouping (p. 313). In the design for this research study, schools were selected as "experimental" and "control". While the formation of these groups was not random, this design assisted in reducing the threat to internal validity from diffusion of treatment that occurs when students communicate about the intervention. This design did pose an external validity threat due to the relatively small number of participants and the specificity of setting. Results, therefore, may not be generalizable to a larger population.

The null hypothesis, H_0: $\mu_I=\mu_c$ stated that there was no significant difference between the groups. This study tested the research hypothesis, H_A: $\mu_I>\mu_c$ to determine whether the intervention was significant.

Research Methods

This quasi-experimental design required five types of data collection: demographic data for school sites, document review of student assessment reports, qualitative coding of student responses during coaching, aggregate pre and post intervention survey scores, and post-intervention student questionnaires. Quantitative data included descriptive statistics of the control and intervention sites, analysis of pre and post SOS scores, and exit questionnaires. During the second phase of the intervention protocol, students individually reviewed their assessment profiles and participated in reflective conversations and goal setting. At that time, qualitative field notes were collected and analyzed.

Stages of Data Collection. Data collection took place in three distinct and sequential stages. The first stage occurred before the intervention when demographic and baseline data was collected and subjects were recruited. The second stage was comprised

of collecting qualitative data during individual coaching sessions with students, and quantitative SOS post-testing. The final stage of the study collected student feedback regarding the intervention and was conducted following the completion of all assessments.

Demographic and baseline data. Stage one began immediately following the IRB approval process. Students from the experimental site were invited to participate in the study during an informational session. Through an agreement with the school district, the initial SOS was given to all students in conjunction with the November 4Sight Benchmark assessment. This preexisting data was mined following IRB approval to determine baseline aggregates of effort and importance scores for each site. The school district required 4Sight testing at least three times each school year. These tests were designed to assess student progress toward end of year standards, and are low-stakes tests for students, since there were not graded. Instructions describing the use of the tests were read to students before the administration of each 4Sight assessment as a standard practice in administering the test. During this stage, the researcher met with approximately eighty middle school students at the intervention site to explain the study and to request parent permission.

Data coaching intervention. Once subjects had been recruited, data coaching began. The intervention timeline was amended due to a delay in IRB approval over the winter holiday. In order to keep the interventions aligned with the school assessment dates, the whole-group presentations were delivered between the individual coaching sessions. This change meant that discussions regarding assessment profiles and

assessment types were delivered in the first individual coaching session rather than in the whole group presentation.

In January, each student met with the researcher to review their results from the 4Sight test given in November. Students also reviewed their assessment profile over two years. This profile included 4Sight and PSSA results. Three students were new to the school district and did not have any data in their profile from previous years. Two other students were missing 4Sight results from November, although the students recalled taking the assessment. During those sessions students discussed their feelings about their profiles and their most recent assessment, estimated the effort they gave on the November test, identified influential factors that contributed to their achievement, and set personal goals. Some students identified behavioral goals while others set achievement goals or actual scores that they thought sought to attain.

In February, students participated in a whole-group discussion designed to establish common language and understanding around the subjects of assessment, data, and effort. This presentation was originally planned to take place in two sessions, but was completed for each grade in a single hour-long session since some of the information had been discussed with students in January. The presentation, which was interactive in nature, provided general information on data collection, the differences between growth and achievement, and how to read a student assessment profile. Students also learned about the relationship between effort and achievement, research findings to support the concept, and practical ways that they could recognize and increase their effort in learning and in testing. This session was delivered in a whole-group lecture and discussion format lasting approximately 60 minutes. Two sessions were provided, one for seventh grade

students and one for eighth grade students. The content of this presentation was focused on assessments, data, and effort. Mastery of these ideas was not expected as they were reinforced and revisited during individual coaching sessions. The following were objectives and key ideas for the whole group session. Slides from the presentation can be found in Appendix C.

1. What is data?
    - Types and uses
    - How its collected (Performance Tracker)
    - Proficiency levels
    - How it affects students (like a credit report)
2. What is effort?
    - What it looks like in class
    - What it looks like on tests
    - What does research tell us about effort?
    - How we can improve effort

Student responses to those sessions were mixed. Most students were passive and respectful, while a few were very engaged. In the seventh grade session, one student had to be excused for continued disruption. Students in the eighth grade session were overall more interest in discussing motivation and test effort. It was noted that the whole group presentation would likely not be effective as a stand-alone intervention. Student attention waned over time, and some students appeared to be disengaged. The usefulness in presenting this material was that it provided a point of reference for later coaching sessions.

The second part of the intervention consisted of individual meetings where students reviewed their assessment profiles, reflected on their feelings about the data, set goals for the next round, and discussed test engagement. This feedback protocol was designed based on the findings of Bangert-Drowns, et al.(1991) regarding the importance

of mediating "mindfulness," and on research regarding the connection between and self-regulation and achievement motivation (Hattie & Timperley, 2007). Feedback procedures were influenced by findings of the latter study, which included reflection, goal setting, and a review of data to measure progress.

In late February, following the whole group presentations, students received their second individual coaching sessions and received feedback on the 4Sight assessment taken earlier that month. During those coaching sessions, students reviewed their most recent 4Sight tests, compared their scores with the November baseline assessment, reflected on their goals, and, in some cases, made connections to the whole group presentation. Students discussed factors that either limited or enhanced their achievement and engagement. Students were again asked to estimate their effort on this test and compare it with their November estimate. All coaching sessions lasted between fifteen and forty minutes depending on how much students wanted to reflect and share. No content interventions were employed during these sessions.

The third 4Sight test was administered to all students in March at School A and School B. Students took the SOS as a posttest assessment at that time.

Exit questionnaires. The third stage of data collection consisted of exit questionnaires given to students who participated in the intervention. The questionnaire asked students to rate their agreement with 11 statements related to test effort, test importance, and recent change in beliefs. Students also rated each aspect of the intervention on a helpfulness scale and provided written opinions on four topics: 4Sight, PSSA, data coaching, and test effort. Figure 4 shows a timeline for this study.

Timeline	Procedure	Tool
November 2011	First 4Sight assessment and Pretest	4Sight and SOS
November 2011	Whole group discussion	PowerPoint presentation
December 2011	First intervention period	Field notes
January 2012	Second 4Sight assessment	4Sight
February 2012	Second intervention period	Field notes
March 2012	Third 4Sight assessment and posttest	4Sight and SOS
March 2012	Exit questionnaire	Exit questionnaire

Figure 4. Timeline for the research study

Description of Instruments. Four instruments were used to collect qualitative and quantitative data. The methods for analyzing the data and a descriptions of each tool are provided.

4Sight.

Instrument description. 4Sight tests are criterion-referenced, benchmark tests that have been designed to project student proficiency on the PSSA (Pennsylvania Training and Technical Assistance Network, 2011). They are administered periodically through the year to assist teachers in analyzing student achievement and mirror the format and content of the PSSA. These tests are designed to be administered either as a web-based assessment or via paper and pencil. Both schools in this study selected paper-pencil assessments.

Assessments were comprised of multiple choice and open ended items, and were constructed to assess student progress toward end-of-year standards. Students took a

reading and a math assessment in separate sessions lasting approximately one hour each and were not timed. Multiple choice items were scanned and scored, while open-ended items were hand-scored by classroom teachers using scoring criteria to assign point values ranging from zero to four points. Each test had approximately 50 multiple choice items and three open ended items.

Data collection. Benchmark data were collected three times during the school year during predetermined assessment dates. The tests were given in reading and math, machine scored, and uploaded to assessment collection software. The results were reviewed with students, but were not collected or analyzed for the purpose of this study.

Data analysis. The data collected from these assessments were used to help students in the intervention group reflect on mastery of their goals and to reflect on their motivation and engagement in class work and on assessments. Scores were charted for each student as a measure of their motivational increase or decrease. School personnel analyzed these data for lesson adjustments, and some choose to give students feedback on content. This process was completely independent from the use of data in this study. Student achievement data was not analyzed in this study as the research questions were related to effort and importance, not learning. Because there are students who exert maximum effort, but do not achieve for a variety of reasons, it cannot be assumed that all students would achieve more if they would only try harder. Since it cannot be known merely by looking at student test results whether low scores are attributable to lack of effort or to lack of knowledge, these constructs must be considered separately. This study aimed to maximize aggregate effort—whether this had an effect on achievement is a question for further research.

Student Opinion Scale.

Instrument description. The SOS (Sundre, 2007) provided quantitative, self reporting of student motivation on low-stakes tests. This survey was designed to be administered following a low-stakes test and consisted of a self-assessment on 10 items that measured two factors: effort and importance.

The first factor was assessed by five questions to determine the level of effort and persistence that a student gives during the exam (Appendix A). An example question from this section was, "I engaged in good effort throughout this test." The second factor was measured by five different questions to determine the personal relevance or importance of the test. An example question from this section was, "I would like to know how well I did on this test." Examinees rated their agreement using a five point Likert scale. Scores were totaled for each subscale resulting in a score range from 5 to 25. Students typically took this survey immediately following task completion.

The SOS has been administered to over 15,000 students, mostly during low-stakes tests at James Madison University. Several studies have been conducted to assess reliability and validity of this instrument (Thelk, Sundre, Horst, & Finney, 2009; Wise, 2006) and they have generally found the SOS to be valid in measuring student effort and importance on low-stakes tests. The reliability evidence for the total is between .80 and .89 and is consistent even when subsections are used in isolation. Variability of scores is reduced in situations where student orientation is high-stakes. In these situations, students are inclined to consider the test important and therefore, generally exert more effort (Sundre, 1999).

Data collection. The SOS was given using the administration guidelines developed with the survey. Teachers at each site were trained prior to the first data collection period. One teacher from the School A and two teachers from School B conducted all SOS administration. In all cases, the SOS was administered in conjunction with the 4Sight reading test, although some students took the math section at the same time. The results of all SOS tests remained anonymous. All students were coded by a unique combination of letters and numbers derived from their homeroom teacher's name, the day of the month that they were born, the first letter of their middle name, and their mother's first initial. Students at the control site received similar codes so that their pre and post scores could be matched.

Data analysis. Mean scores for the effort and importance scales of the SOS were analyzed to determine the group's general effort and sense of importance on the benchmark assessment. Descriptive statistics were calculated on pretest and posttest scores including range, mean, and standard deviation. Distributions for effort and importance scales on pretest and posttest were considered for normality at both sites. Pre-intervention and post-intervention mean scores were analyzed within and between groups using an analysis of variance (ANOVA) to determine the significance of the intervention.

During the data coaching sessions, student ability was considered as high, low or erratic by using assessment profiles for 4Sight and PSSA over the previous two years. This was a fairly easy task, as each test was color coded based on proficiency with advanced scores showing blue, proficient scores as green, basic scores as yellow, and below basic scores as red. Scores that were all or predominantly blue and green were

identified as “high.” Students with all or predominantly yellow and red scores were identified as “low.” Scores that had a variety of colors were identified as “erratic.” Students who were new to the district or had incomplete data were identified as “not known.”

Because student ability beliefs may have affected effort, t-tests of mean differences in pretest and posttest scores for effort and importance were run. These tests compared each assessment profile group with the remaining members of the experimental sample to determine whether changes were more significant for an ability subgroup.

Field notes.

Instrument description. Field notes were collected during individual student coaching sessions. Student responses during the intervention were collected using a standardized form that detailed the components of individual coaching. Participants reflected on their feelings about their performance, reviewed their performance consistency over time, discussed possible changes, and set goals for future test effort and performance. A sample of the form used for collecting field notes during data coaching sessions can be found in Appendix D.

Data collection. Student responses that were collected on the field note form were transcribed to Microsoft Excel in fields that reflected the standard protocol of the interviews. The notes were coded for positive and negative effort and importance, and for factors that contributed to those attitudes. Transcription occurred following each coaching session. Data was disaggregated by achievement profile, comparing those with consistently high or low performance against students who had an erratic performance history. Students with erratic profiles offered particular insight in that they had

demonstrated the ability to achieve but had not sustained a high level of achievement over time. Statements remained anonymous and were collected only to further explain student views regarding test engagement, motivation, and assessment task value.

Exit questionnaires

Instrument description. Exit questionnaires were given in the final stage of the study to determine what specific elements of the intervention students found most helpful. Students rated the following aspects of the intervention: assessment knowledge, personal achievement feedback, goal setting, and reflection. This questionnaire consisted of 11 statements with which students rated agreement using a five point Likert scale. Four questions assessed test importance and seven assessed general effort. Additionally, three questions asked students to assess change in effort or importance and two questions each assessed parental or school staff influence. Students also rated each aspect of the intervention on a helpfulness scale and provided opinions on four intervention topics: PSSA, 4Sight, data coaching, and test effort.

The questionnaire was field-tested using a focus group of students at a feeder high school. The group participated in a similar intervention during the prior year as eighth graders. They were not included in this study; however, their feedback was instructive in improving the language and purpose of the instrument.

Data collection. Exit questionnaires were collected during the final intervention phase with students. Questionnaires were administered by volunteer teachers in a whole group setting. Students were reminded of the voluntary nature of this phase of the study.

Data analysis. Descriptive statistics were reported from exit questionnaires. Tables that showed the percentage of student agreement for each statement were

constructed. The data from the questionnaire was instructive in determining student views about the data coaching experience: it will be useful in future interventions.

Procedures

Quantitative data. In a pretest - posttest design, students at School A and School B took the SOS in November, providing baseline test scores for all students at both sites. Student agreement with 10 statements were measured on a one to five scale, where 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, and 5=Strongly Agree. Four statements were worded negatively and were reverse scored. Five statements each determined two subscale scores: one for *importance* and one for *effort*. The score for each subscale was derived by summing the score for each statement. Subscale scores ranged from 5 to 25, with lower scores indicating lower beliefs about importance or effort. These scales were independent of each other so it was possible, for example, for students to have felt that a test was important, but to have a low score on the effort scale. Total scores were not considered in accordance with directions in the administration manual.

Students at School A were invited to participate in the study during meetings that were scheduled during the first few minutes of their social studies class. Students received a letter describing how they could be involved and that parent permission was necessary. Those who returned signed permissions by the end of the week were select and offered compensation of $1 or participation in a pizza party. Parent contacts were made, and additional forms were given to students during the week. As student assents were also required, student willingness to participate was demonstrated by first securing parental permission. Homeroom teachers assisted with student reminders and in

collecting permission forms. Student assents were signed prior to the first data coaching session and student codes were identified so that baseline data could be collected.

Baseline scores for effort and importance were derived from the November pretest for the experimental site participants only. Institutional Review Board approval did not require parental permission or student assent at School B, since student scores were anonymous. Effort and importance scales scores were calculated for assented participants using directions found in the SOS Administration Manual. Since the number of participants was relatively small, tests were hand scored and then double checked for accuracy. These baseline scores were recorded in an Excel spreadsheet using student codes.

Posttest SOS scores were taken in March following the final data coaching session. Students had just completed their third 4Sight Benchmark test and were preparing to take the PSSA the following week. Posttest scores for participants were scored in the same manner as the pretests and were added to the Excel workbook. SOS scores from the control site were selected in March, using only those that had a complete pretest and posttest pair. These were scored similarly and added to the Excel workbook.

Data from Excel was transferred to SPSS, and summary statistics were calculated for effort and importance at the experimental and control sites to compare score distributions and evaluate normality. Graphic representations of these findings were also constructed. School A had outliers in all data samples and some skew in the pretest scores for importance, but the deviations were not severe enough to warrant any transformations.

A repeated measures ANOVA was run for both effort and importance to consider whether the effect of time was significant for each group. This test included a time effect, a group effect, and the interaction. Marginal means were calculated from the data and plotted to provide graphical representation for a main effect or interaction. Additional analyses were also conducted in an attempt to find significance. First, the data set was altered to eliminate greatest outlier and then again with the three largest outliers. ANOVAs were repeated on the data without the outliers. Next, a t-test for independent samples was run using the differences between pretest and posttest values as the dependent variable. Subtracting the post score from the pre score resulted in a single difference variable testing the hypothesis that changes were larger for the experimental site compared to the control site. Box's M and Levine's univariate test of equality were also used to evaluate variance between groups.

To consider whether student ability beliefs confounded the findings, SOS data was also analyzed by assessment profile. The change in scores for importance and effort were calculated by finding the difference between the pre and post tests for students at the experimental site only. These scores were disaggregated by assessment profile, and descriptive statistics were calculated on the average change. The mean, standard deviation, and standard error of the mean for each disaggregated group were reported against the remaining cases in the sample. For example, members of the high group were compared against combined low, erratic, and unknown profiles. These numbers were also evaluated with t-tests.

Data from exit questionnaires was also compiled in Excel. There were three sections in this instrument. The first section consisted of 11 Likert-type statements

regarding test effort and importance. Scores for each of the statements in this section were calculated by counting the number of respondents who selected each degree of agreement. Totals were then multiplied by ascribed weight, mirroring the SOS scale, 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, and 5=Strongly Agree. Questions were ranked for strength of agreement using weighted totals.

The second section consisted of five aspects of the data coaching experience. Students were asked to rate these on a six point helpfulness scale. In a like manner, weighted means were calculated with descriptive statistics and agreement rank was determined from the weighted mean. The final section was to have qualitative data as students were given an opportunity to give final feedback in four categories. During a pilot study of this instrument, this section yielded a significant quantity of data. During the actual study, few students wrote anything and all, perhaps recognizing that what they had to say had already been collected during individual data coaching sessions. The few comments that were included only reiterated information previously collected, therefore, no data from the final section of this instrument was analyzed.

Qualitative data. Following each data coaching session, qualitative student responses were entered into Excel. Student information was separated using rows for students and columns for categories. Initial coding of comments was limited to positive and negative comments only. After the second coaching, session all comments were coded and analyzed more completely. At that time, data fields were sorted by assessment profile. Reports were generated by categories resulting in four reports for each category: those for high, low, erratic, and unknown profiles. Category reports were generated for feelings, effort, and importance. Using an open-code model, a code book was created for

attributions for positive or negative effort and importance. Once all reports were coded, summaries of the codes were tallied in Excel, and the data transformed into graphical representations. Codes are found in Appendix E.

Ethical Considerations

This study was designed to examine the effect of data coaching on achievement outcomes. One aspect of coaching involved building relationships with students and therefore, required that the researcher had prior knowledge of and relationship with study participants. To address ethical concerns related to coercion, volunteer teachers acted as liaisons with students. They answered questions and provided space and time for intervention activities.

While all students participated in 4Sight testing, only students from School A could participate in the intervention portion of the study. Students from other schools in the school district did not have the opportunity to benefit from the intervention. To address this concern, principals in the school district will be briefed on the results of this study. They may choose to implement this intervention at their discretion in subsequent years. Since this study was designed to improve effort and not achievement on low-stakes tests, students who chose not participate had the same learning opportunities as participants. The purpose of 4Sight benchmark assessments is to give teachers better information about student achievement so that they can modify their instruction—all students benefitted from this practice. Data coaching experiences were arranged to have minimal impact on student instructional time. All sessions were conducted during the school day and in consultation with supervising teachers.

During the intervention protocol, individual student assessment profiles were shared with students, but data were not collected or analyzed. The school district used a data management system, which provided detailed assessment data for each student and across several years. This information was printed to share with students, but the information remained at the school.

Intervention subjects had the opportunity to complete an exit questionnaire following the final phase of the study. Although permission for the questionnaires was secured as part of consent for participation, students' participation in this phase of the study was voluntary and was considered independent of their participation in the intervention.

This research proposal was approved by Drexel University's Institutional Review Board in December 2012. All aspects of this study were exempt under category 1 (research involving normal educational practices) and category 2 (educational test data and surveys). As part of the assent process, students were counseled that they had the right to withdraw from the research at any time.

Chapter 4: Findings and Results

The purpose of this mixed-methods, quasi-experimental study was to determine the effect that a data coaching intervention had on student reported effort and importance scores on the SOS (Sundre, 2007). Field notes were recorded, coded, and analyzed to enhance SOS results. These data examined student feelings about their assessment profiles, their beliefs about test importance, and their reasons for exerting, or failing to exert, effort. Exit questionnaires were collected to find what aspects of the interventions students found most helpful, and to discover what effects the process had on student attitudes toward low-stakes tests.

Participant Demographics

Thirty-six students gained parent permission and provided assent at the experimental site, a response rate of 45%. Of these, ten were male (n = 10) and twenty-six were female (n = 26). Participants were equally divided between the seventh and eighth grades at eighteen participants each. Two (n = 2) students were classified as English Language Learners, but had attended school in the United States for more than a year. Seven students (n = 7) received special education services.

Scores from School B were paired using student codes. Students were eliminated if they did not have both pretest and posttest scores either due to absence, or failing to provide consistent code information. A sample of student scores equal to the number of experimental participants (n = 31) was randomly selected using the fourth score from each classroom when the surveys were ordered by date of birth. This yielded 19 seventh grade and 12 eighth grade scores for the control site. All students participated in the SOS at the control site, and a sample equal in number to the experimental sample was taken

randomly from among the same grades as the experimental site. Student identifiers were not used at the control site, and demographic information was not available.

Findings

Student Opinion Scale. Although 36 students participated in the study, 5 did not have matching pretest and posttest SOS scores. Therefore, 31 samples were selected from the control group, yielding a sample size of 62 (n = 62). SOS scores were calculated by summing student agreement on five effort statements ("I gave good effort throughout this test.") and five importance statements (Doing well on this test was important to me."). Although score interpretation is relative, using a five-point Likert-type scale, a score of 20 for each construct would indicate general agreement. For the purposes of this study, score increases or decreases were used to interpret student beliefs about low-stakes tests over the course of the intervention.

First, means, standard deviations, and the range of scores were calculated for both the experimental and the control group. In the experimental group, the pretest scores for the importance scale ranged from 10 to 25 with a mean of 19.32 and a standard deviation of 3.124. The posttest importance range was from 5 to 25 with a mean of 18.29 and a standard deviation of 4.713. The pretest scores for effort scale ranged from 7 to 24 in the experiment group with an average of 18.58 and a standard deviation of 3.529. The posttest scores for effort ranged from 5 to 23 with a mean of 16.61 and a standard deviation of 4.425.

For the control group, the pretest scores on the importance scale ranged from 12 to 25 with a mean of 18.81 and a standard deviation of 3.616. The posttest range for importance was from 12 to 25 with a mean of 18.58 and a standard deviation of 3.686.

The pretest score for effort ranged from 11 to 24 with an average of 18.16 and a standard deviation of 3.616. The posttest scores ranged from 11 to 25 with a mean of 17.45 and a standard deviation of 3.443. Table 2 shows summary statistics for the pretest and posttest SOS by treatment group and the main variables.

Table 2. *Summary Statistics for Pretest and Posttest Student Opinion Scale*

	n	Minimum	Maximum	M	SD
Experiment					
Importance- Pre	31	10	25	19.32	3.124
Effort-Pre	31	7	24	18.58	3.529
Importance-Post	31	5	25	18.29	4.713
Effort-Post	31	5	23	16.61	4.425
Control					
Importance- Pre	31	12	25	18.81	3.637
Effort-Pre	31	11	24	18.16	3.616
Importance-Post	31	12	25	18.58	3.686
Effort-Post	31	11	25	17.45	3.443

Second, student score distributions were tested for normality on each subscale, by group and for pre and post test scores. Figure 5 shows a near normal distribution box plots for student scores on importance. There were some outliers in the experimental group, and the pretest results for the experimental group had some skew since the median line was not centered. However, these deviations from normality were not large enough to require alterations to the data. Nonetheless, additional tests were run to ensure that outliers did not impact the findings. These tests are discussed further on page 83.

The box plots for effort are similar to those for importance. Overall, the distributions were close to normal, especially for the control group. There were only a few outliers and some skew in the pretest scores for the experiment group. Because the

sample size was relatively small and dropping these scores would also drop information, no changes were made to these variables for the initial analysis. The assumptions underlying the repeated measures ANOVA appeared to have been met in that the box plots indicated distributions that were close to normal.

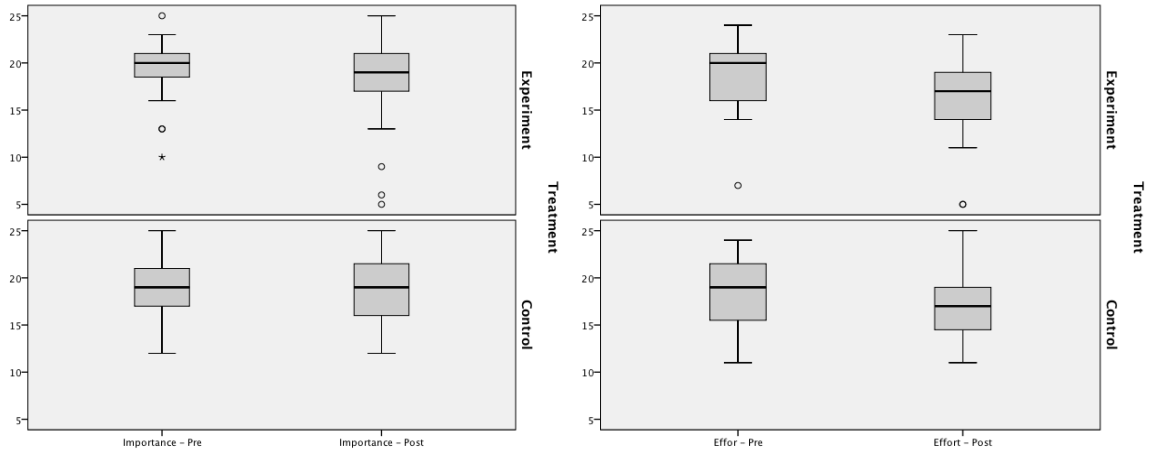


Figure 5. Box plots of normality for the importance and effort subscale, pretest and posttest by group

Third, a repeated-measures ANOVA was carried out to determine if there was a significant change between and among experimental and control groups using pretest and posttest scores for both effort and importance. In this model, time was treated as a within-subjects factor, and the group was a between subjects factor. Table 3 summarized the F-tests for importance while Figure 6 graphed the marginal means of the importance variable.

Table 3. *Repeated Measures ANOVA for the Importance Subscale*

	Sum of Squares	df	Mean Square	F	Sig	Partial Eta Squared	Observed Power
Time	12.266	1	12.266	1.357	.249	.022	.209
Group	.395	1	.395	.019	.890	.000	.052
Time X Group	5.040	1	5.040	.558	.458	.009	.114

The graph indicated an interaction since there was a larger drop from pretest to posttest among the experimental group compared to the control group. However, if students ascribed greater importance to the test, scores would have been expected to increase. In addition, the interaction did not turn out to be significant $F(1, 60) = .558$, $p > .05$ and none of the direct effects were significant. The overall change for all subjects between time 1 and time 2 was not significant $F(1, 60) = 1.357$, $p > .05$, and neither was the overall average between the two groups $F(1, 60) = .019$, $p > .05$. All the effect sizes were also small. The eta-squared for the main effect of time was only .022; it was less than .001 for the group main effect; and it was .009 for the interaction.

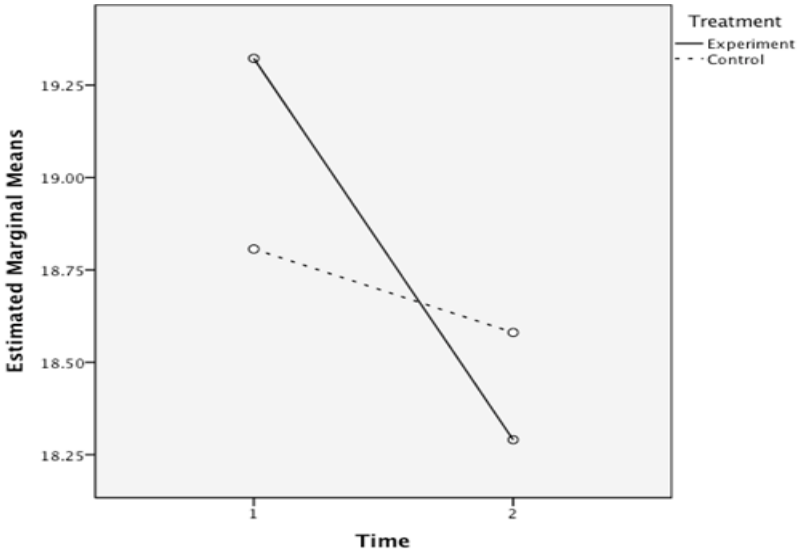


Figure 6: Estimated marginal means of importance

The observed power column in Table 3 indicated that for the observed effect size and a .05 cut-off for significance on a table of critical values, power was quite low (<.21 for all the tests). In addition, it was not possible to reject the null hypothesis that error variances were equal between groups. Box's M is a multivariate test whose null hypothesis is that the error variances and co-variances are equal between the two treatment groups. A non-significant result implies that the assumption is met. The test for importance yielded an insignificant p-value (p = .128). In addition, Levene's univariate test of equality showed that one could not reject the null hypothesis that the pretest scores (p = .205) and posttest scores (p = .524) had equal variances between groups. A much larger sample would have been needed before it would be possible to say that the differences were statistically significant.

An ANOVA was also run for the effort variable with similar results. Table 4 and Figure 7 have provided the results. The plot again indicated there may be an interaction

with the difference between pretest and posttest scores being larger for the experiment group than for the control group; however, the statistical tests did not show a significant result for the interaction $F(1, 60) = .072$, $p > .05$.

Table 4. *Repeated Measures ANOVA for the Effort Subscale*

	Sum of Squares	df	Mean Square	F	Sig	Partial Eta Squared	Observed Power
Time	55.556	1	55.556	5.831	.019	.089	.661
Group	12.266	1	12.266	1.287	.261	.021	.201
Time X Group	1.363	1	1.363	.072	.789	.001	.058

There was a significant effect for the main effect of time $F(1,60) = 5.831$, $p = .019$ in that posttest scores compared to pretest scores tended to be lower on average for all subjects regardless of treatment group. Time was able to explain 8.9% of the total variability in the dependent variable. The main effect for group was not significant $F(1, 60) = 1.287$, $p > .05$ and the observed power for this effect was very small, indicating that a much larger sample size would be required in order to find a significant result. Box's M test for equality of variances and co-variances between groups was not significant ($p = .555$), and Levene's Test was not significant for both the pretest scores ($p = .185$) and the posttest scores ($p = .737$). In both tests the null hypothesis is retained. Overall, there did not appear to be evidence that the pretest and posttest scores changed in a manner that was significantly different between the experiment and the control groups.

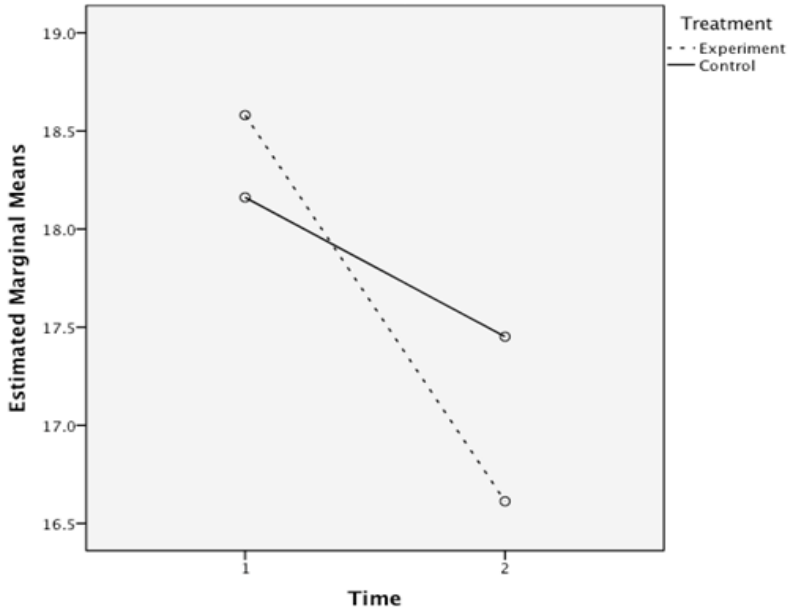


Figure 7. Estimated marginal means of effort.

Fourth, some additional analyses were attempted to determine if outliers were having an influence on the results. In a first analysis, the largest outlier (represented with the "star" in Figure 5) was the only observation dropped. Doing so did not change the non-significance of the results. After removing this single observation, the next three largest outliers were also removed, and the analysis was re-run. Once again, the results were not significant. This was not surprising if the lack of significance was due to power, since dropping observations reduced power even further. These tests did demonstrate that outliers were not contributing to the lack of significance.

Fifth, a simple t-test was run on the differences between pretest and posttest scores. These results are summarized in Table 5. For the importance variable, the average decrease pretest score to posttest score among the experiment group was 1.032 (SD = 4.950); this difference was .226 (SD = 3.413) for the control group. Although there was a

difference between the two groups, it was not statistically significant for importance t(60) = .747, p = .458.

Table 5. *Means of Difference Scores*

Variable	Group	M	SD	T	df	p
Importance Difference	Experiment	1.032	4.950	.747	60	.458
	Control	0.226	3.413			
Effort Difference	Experiment	1.968	4.701	1.135	60	.261
	Control	0.710	4.002			

The experimental group also demonstrated a larger change in effort scores (1.968, SD = 4.701) compared to the control group (.710, SD = 4.002). Again, the t-test is non-significant for effort t(60) = 1.135, p = .261, meaning one cannot reject the null hypothesis of no differences between groups.

Finally, an analysis of the change in scores within the experimental group was considered. In this test student scores were disaggregated by assessment profile type to determine whether student ability beliefs were a confounding variable. Using color coded proficiency levels, all subjects (n = 36) were placed into groups based on their achievement pattern over two years: “high” (n = 7), “low” (n = 11), “erratic” (n = 14), and “not known” (n = 4). However, when SOS scores were paired for pretest and posttest analysis, unpaired samples were eliminated. The resulting t-test groups were “high” (n = 6), “low” (n = 11), “erratic” (n = 12), and “not known” (n = 2).

First, mean differences in pretest and posttest scores from each assessment profile group were compared see if changes were significant for any subgroup. Figure 8 shows the mean change in pretest and posttest scores on the importance scale by assessment profile. For the high group, there was a slight increase from 19.5 (SD = 5.167) to 20 (SD

= 3.286). However, this change was not statistically significant t(5) = -.311, p = .768. The only other category that showed an increase from pretest to posttest was the not known category. Their pretest score was 17 (SD 5.657), while the posttest score was 19 (SD = 2.828). However, with only two observations, the effect could not be called statistically significant t(1) = -.333, p = .795). Meanwhile the low group showed a mean decrease from 19.27 (SD = 2.970) to 18.09 (SD = 5.127). This difference was also not significant t(10) = .836, p = .422. Finally, the erratic group showed a decline from 19.67 (SD = 1.557) to 17.50 (SD = 5.351). The change was not significant t(11) = 1.397, p = .190.

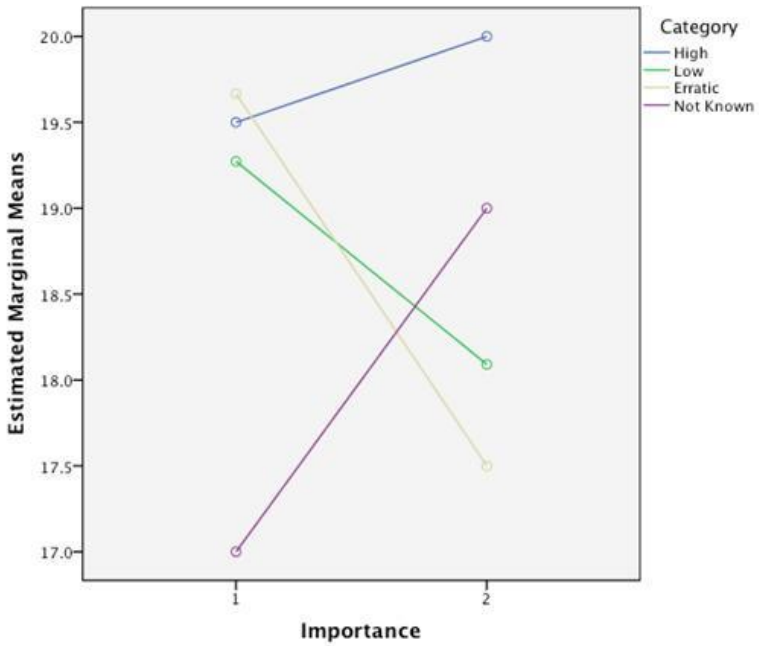


Figure 8. Estimated marginal means of importance within experimental subgroups by assessment profile

Figure 9 shows the same information for the effort scale. Each group showed a decline from pretest to posttest. In the high group, the decline was from 20.5 (SD = 1.517) to 15.33 (SD = 5.715). This was not significant t(5) = 2.289, p = .071. For the low

group, the decline was from 17.55 (SD = 4.547) to 16.45 (SD = 5.007). This is also non-significant t(10) = .860, p = .410. The erratic group showed a decline from 18.08 (SD = 3.029) to 17.08 (SD = 3.450), which was also non-significant t(11) = .797, p = .443. Finally, the two subjects in the unknown group showed a decline from 21.5 (SD = .707) to 18 (SD = 7.071), which was not significant t(1) = .778, p = .579.

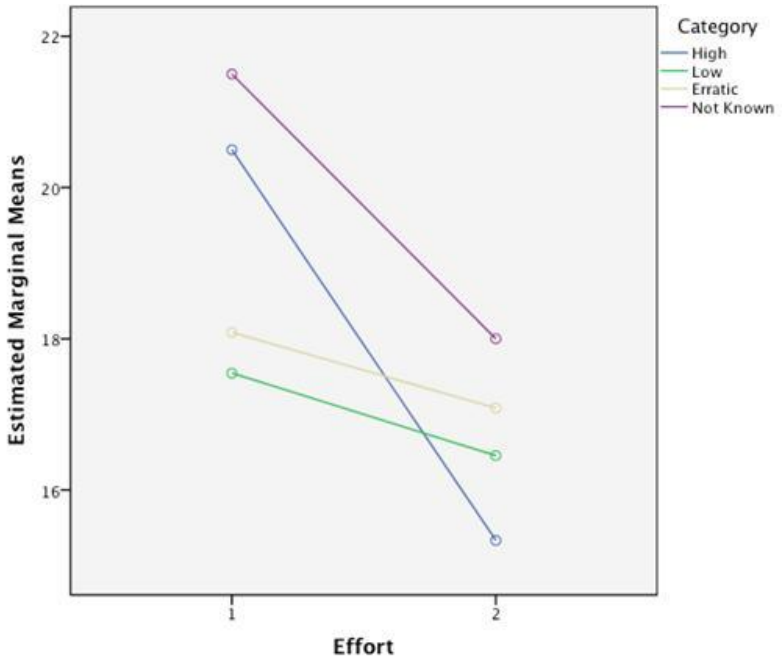


Figure 9. Estimated marginal means of effort within experimental subgroups by assessment profile

Another way to consider the data is to compare the average change in scores between each assessment profile group and all the remaining cases in the experimental sample. For example, members of the high group would be compared against the low, erratic, and unknown profiles. Because the interest was in the amount of change from

pretest to posttest, differences in scores were calculated and these numbers were evaluated with t-tests.

Table 6 shows the average change scores for the importance scale. Change in scores was determined by subtracting the posttest score from the pretest score. The average change for the high group was -.500 (SD = 3.937), meaning that the average posttest score for importance was .500 higher than the average pretest score. This compared to an average decrease of 1.400 (SD = 5.164) for subjects not in this profile. The average change in score for the low group was 1.182 (SD = 4.687), meaning that the pretest score was 1.182 points higher on average compared to the posttest score. This compared to an average change of .950 (SD = 5.206) for subjects not in the low group. Finally, the average change for the erratic profile was 2.167 (SD = 5.347), compared to .316 (SD = 4.667) for everybody else.

Table 6. *Average Change in Importance Scores (Pretest – Posttest)*

Profile	n	M	SD	Std. Error Mean
Not High	25	1.400	5.164	1.033
High	6	-0.500	3.937	1.607
Not Low	20	0.950	5.206	1.164
Low	11	1.182	4.687	1.413
Not Erratic	19	0.316	4.667	1.071
Erratic	12	2.167	5.374	1.551

Table 7 shows t-tests for each comparison. In every case, the null hypothesis of equal variance cannot be rejected (High: F = .235, p = .631; Low: F = .221, p = .641; Erratic: F = .192, p = .664). The difference between the average high and not-high

profiles was 1.900 (SE = 2.261), which was not significant t(29) = .840, p = .408. The difference between the average low and not low profiles was -.232 (SE = 1.889), which was also not significant t(29) = -.123, p = .903. The difference between the average erratic and not erratic profiles was -1.851 (SE = 1.824), which was, again, non-significant t(29) = -1.015, p = .319.

Table 7. *T-tests of Mean Differences in Importance Change Scores*

	Equal Variance						
	F	P	Diff	SE Diff	t	df	p
High	0.235	0.631	1.900	2.261	0.840	29	0.408
Low	0.221	0.641	-0.232	1.889	-0.123	29	0.903
Erratic	0.192	0.664	-1.851	1.824	-1.015	29	0.319

The same tests were run for effort. Table 8 shows the average change in scores on the effort scale. The average change for the high group decreased from pretest to posttest by 5.167 (SD = 5.529), compared to a change of 1.200 (4.252) for the non-high profile group. The average change for the low group was decreased from pretest to posttest by 1.000 (SD = 4.171), compared to a change of 2.500 (SD = 4.989) for the non-low group. The average change for the erratic group was to decrease from pretest to posttest by 1.000 (SD = 4.348), compared to a change of 2.579 (SD = 4.925) for the non-erratic group.

Table 8. *Average Change in Effort Scores (Pretest – Posttest)*

High	n	M	SD	Std. Error Mean
Not High	25	1.200	4.252	0.850
High	6	5.167	5.529	2.257
Not Low	20	2.500	4.989	1.116
Low	11	1.000	4.171	1.258
Not Erratic	19	2.579	4.925	1.130
Erratic	12	1.000	4.348	1.255

Table 9 shows the results of t-tests comparing the difference in effort scores between groups. The null hypothesis of equal variances holds in all three cases (High: F = .020, p = .888; Low: F = .295, p = .591; Erratic: F = .090, p = .767). The difference in average scores between high and not high profiles was -3.967 (SE = 2.045), which was not significant t(29) = -1.940, p = .062. The difference in average scores between low and not low was 1.500 (SE = 1.773), which was also not significant t(29)= .846, p = .404. The difference in average scores between erratic and not erratic was 1.579 (SE = 1.738), which was also not significant t(29)= .908, p = .371.

Table 9. *T-tests of Mean Differences in Effort Change Scores*

	Equal Variance						
	F	p	Diff	SE Diff	t	df	p
High	0.020	0.888	-3.967	2.045	-1.940	29	0.062
Low	0.295	0.591	1.500	1.773	0.846	29	0.404
Erratic	0.090	0.767	1.579	1.738	0.908	29	0.371

All statistical tests showed a lack of significance in SOS scores between and among experimental and control groups using pretest and posttest measurements for both

effort and importance. An additional lack of significance was found in comparing pretest and posttest scores and differences between pretest and posttest scores by assessment profile type within the experimental group.

Field notes. Field notes were collected during data coaching sessions using a form to standardize conversations as much as possible. The first session focused more heavily on reviewing past student assessment scores and on discussing influences that contributed to these scores than did the second session. During the first session, students also discussed their assessment profiles and their feelings about their performance over the past two years. Because the second coaching session occurred within days of 4Sight administration, students were more specific about content questions as they reflected on their performance. Although student achievement was not analyzed as part of this study, it is noteworthy that student performance was significantly lower on the math portion of the January 4Sight than on the one taken in November; many students reflected on the reason for this change during their second interviews. In all coaching sessions student statements were recorded on paper and participants were asked to review the notes for accuracy at the conclusion of each session.

Because students discussed test effort in considering the most recent tests that they had taken, influences on their effort fluctuated over time. Effort comments were therefore coded according to influence and were disaggregated by coaching session and by test profile. Statements related to test importance were coded in a similar way. However, the majority of comments about test importance occurred in the initial coaching session, making analysis over time less meaningful.

Feelings. Students identified positive, negative and mixed or neutral feelings when viewing their assessment profiles or individual test scores. A list of words associated with feelings was made available for students, but was only used on two occasions. Feelings comments were sorted into positive, negative and mixed or neutral categories by profile pattern. Within each category comments were coded for specific feeling. Positive feelings included proud, smart, improving, successful or confident, positively surprised, and general positive feelings of good or happy. Negative feelings included those for doubt or failure, worry or nervousness, disappointment, regret, embarrassment, anger, overwhelmed, frustration, negatively surprised, and general negative feelings of bad or sad. Mixed or neutral codes included feelings of confusion, acceptance, or balanced responses that considered both positive and negative feelings. One student declined to identify any feeling words. Physical responses were noted when they exceeded common smiling or frowning. One positive physical response was noted as "big smile" while four negative physical responses were all noted as crying or weeping. Analysis shows interesting patterns among assessment profile types that may explain student preferences. Table 10 summarizes student feelings by achievement pattern.

Table 10. *Student Feelings Regarding Assessment Profile by Achievement Pattern*

	High (n = 7)	Low (n = 11)	Erratic (n = 14)	Not Known (n = 4)
Positive	9	0	10	1
Negative	9	23	12	7
Neutral or Mixed	4	5	7	0

Students with high profiles shared equally in positive and negative comments. Four students said that they were proud while two students each said that they felt smart or successful. One comment was attributed to a general feeling of "happy." All the negative comments (n = 9) were attributed to worry, regret, anger or surprise. In every case, these students either liked seeing their results or discussed that they wanted to improve.

Of the 23 feelings for students with a low assessment profile, none was positive. Students reported disappointment (n = 5), regret (n = 3), frustration (n = 2), embarrassment or shame (n = 2), anger (n = 1) and worry (n = 1). Five students said that their scores made them feel "bad," and four students cried. Students with an erratic profile had a wider range of reactions. One or two comments were counted for every positive feeling (n = 10), and one or two negative comments were made for every negative feeling (n = 12), except disappointment and embarrassment. None of the erratic students described those feelings.

Importance. Students made a total of 70 comments related to the importance of tests. Ten of those comments related to the PSSA. Eight students spoke specifically about the test as having importance. Six of these comments compared the PSSA to the 4Sight, ascribing greater importance to the former. An example of this feeling is from a student with an erratic assessment profile, "I don't like 4Sights. There's no point to it. I find it pointless, but I still try. The PSSA is a state test and it counts." Another student described the importance of 4Sights more to teachers than to self, "4Sights are only important to get me ready for PSSAs; it's more review. It's important to teachers so they

know what to teach." One student with an erratic pattern had the opposite view, "4Sight tests are important, but not PSSA. I think PSSAs are placement tests."

Sixty remaining comments pertained specifically to 4Sight testing. These comments were coded as positive (those that indicated student ascribed importance) and negative (comments indicating a lack of importance). These comments were tallied according to assessment profile. Table 11 summarizes student comments for importance. Student beliefs about 4Sight test importance varied, although students with high assessment profiles tended to ascribe more importance than students with low or erratic test patterns. Overall, students with low and erratic assessment profiles had the most negative comments, and high students almost always viewed tests as important.

Table 11. *Student Comments for Importance by Assessment Profile*

	High (n = 7)	Low (n = 11)	Erratic (n = 14)	Not Known (n = 4)
Positive	11	12	14	2
Negative	1	11	9	0

In addition to coding comments about test importance as positive or negative, student attributions for these beliefs were also analyzed when given. Of 40 comments recorded, attributions for positive beliefs were coded for parent influence (n = 2), teacher or school influence (n = 2), feedback (n = 4), awareness of consequence (n = 9), preparation for higher stakes tests (n = 8), and a belief that testing is an integral part of learning (n = 10). Figure 10 shows the distribution of these comments. The greatest number of student comments acknowledged the importance of benchmark testing in the

learning process. One student shared, “I think tests are important. If you don’t take tests, there’s no point in learning.” Three students recognized that tests were useful in their desire to know how well they had mastered material. Two students said that tests could tell if they had paid attention.

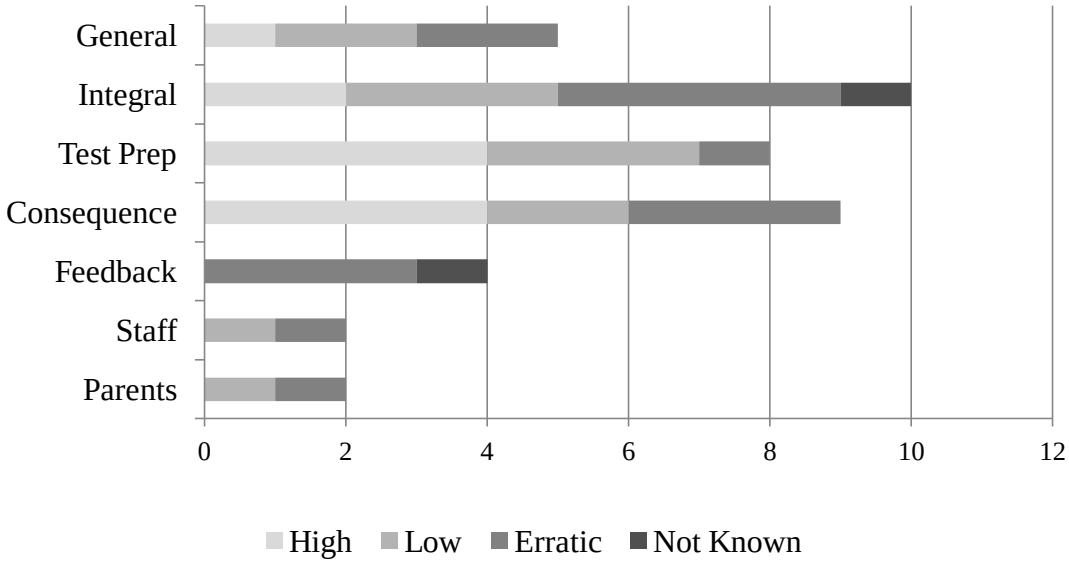


Figure 10. Attributions for positive test importance disaggregated by assessment profile type

Consequences also influenced student beliefs. All positive consequence comments were made by eighth grade students. Seven of those were in the midst of applying for admission to high schools and recognized that these assessments were considered as part of the application process. One student said, “…I just had an interview with SciTech High and I know now that they look at things like this.” Another student viewed test scores as important in determining whether students would be promoted to high school. The final comment related to the importance of test scores when applying to

college. Given the pressure on teachers and students to perform on high stakes tests, it was not surprising that eight students recognized the 4Sight assessments as preparation for PSSA tests.

Twenty student responses indicated lack of importance for 4Sight tests. When students gave reasons, they attributed those beliefs to staff, lack of feedback, and lack of consequence. In one case, a student with a low assessment profile attributed negative importance to both staff and lack of consequence: "Teachers tell us the 4Sights don't count. I would try harder if it was part of our grade." Most of the general comments were due to students' beliefs in overuse of assessment: "All these tests are unnecessary, they are overrated. All these benchmark tests are unnecessary. Maybe [we should] take 4Sight at the beginning of the year and PSSA at the end." Students also felt that lack of feedback conveyed lack of importance: "No one really talks to me about these tests…." Figure 11 summarizes negative student importance attribution.

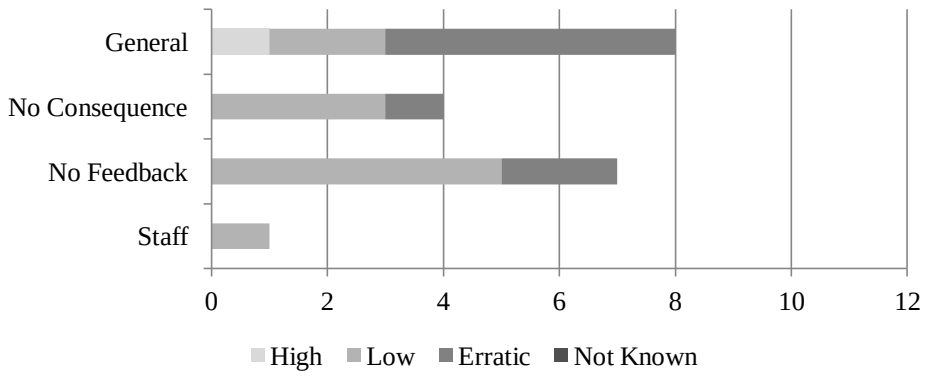


Figure 11. Attributions for negative test importance disaggregated by assessment profile type

Effort. Student responses that described test effort were coded as positive (those that show effort) or negative (those that describe giving up or not attending) and then coded for influence. One comment was coded as neutral. Examples of positive test effort included, "I highlight questions that I have on the test and look back at the articles on the reading portion," and "I tried harder because I saw my scores the last time…and I knew that I had to do my best." Examples of negative effort included, "I'm starting to…give up without trying because it's hard," and "I don't really put effort towards it (4Sight). I don't believe that it actually counts on my grade." Unlike comments for importance that were consistent over time, student comments regarding effort changed over time. For this reason, data was disaggregated by assessment profile and by coaching session. Table 12 summarizes positive and negative comments. Positive and Negative 1 comments occurred during the first coaching session. Positive and Negative 2 comments were noted in the second coaching session.

Table 12. *Positive Student Comments for Effort by Assessment Profile and Coaching Session*

	High (n = 7)	Low (n = 11)	Erratic (n = 14)	Not Known (n = 4)
Positive 1	5	2	12	2
Positive 2	4	3	15	4
Negative 1	9	10	24	0
Negative 2	7	8	10	4

Reasons given in either session for effort-striving included: parental expectations or support, peer support or competition, internal motivation or avoiding consequence,

teacher preparation or support, or strong ability beliefs. Negative influences on effort included: peer distraction, personal lack of focus, lack of teacher support, ability beliefs, cost in time or energy, attitude, and lack of consequence. Some positive and negative comments were general: "I tried my best" or "I didn't try" and were coded as "not attributed". Table 13 shows total positive attributions for effort by assessment profile reported by coaching session.

Table 13. *Attributions for Positive Effort Disaggregated by Assessment Profile and Session*

	High		Low		Erratic		Not Known		
Session	1	2	1	2	1	2	1	2	Total Attribute
Parents	1	0	0	0	3	0	0	0	4
Classmates	1	0	0	0	0	0	0	1	2
Consequence	1	0	1	1	0	2	0	1	6
Teacher	1	2	1	1	7	10	1	2	25
Ability beliefs	1	0	0	0	1	2	0	0	4
Not attributed	0	2	0	1	1	1	1	0	6

Among all positive attributions for effort, teacher influence was cited most often—10 times during the first session and 15 times in the second session. In all cases, students indicated that teacher concern, effective teaching, or test strategy positively affected their test effort. One student described this influence, "Mr. X taught us an easy way to do the open ended [questions] on 4Sight and PSSA." This strategy was similarly

cited and described by other students, “I use the ‘TAG it a 3’ strategy for answering open ended questions. I didn’t know that strategy until this year” and “I give more effort in reading because of the way Mr. X teaches us. It makes me feel smarter than I really am.” Seventeen of the 25 comments attributed to teachers were made by students with erratic profiles.

Table 14 summarizes negative comments by assessment profile across time. Among negative attributions for effort, students from all profiles most often cited the cost of 4Sight assessments in time and energy. Some students admitted complete disengagement, “I don’t like tests; it [*sic*] makes me tired so I go to sleep.” Other students describe guessing practices, “When I’m filling in the bubbles I do a little race to see which letter will have the most [answers]…” and “…[tests] are too much work. I get bored and circle any answer.” When asked if low scores were an indication of intelligence, one student said, “No, I was just being lazy…the test took so long I didn’t want to finish it. I just did anything.” Students also recognized that effort is situational, “Sometimes I put all my effort into it. Sometimes I just give up on the test,” and “On the reading test, the first story I try really hard, but I get tired of reading, so I don’t try as hard.” Students cited open ended questions as higher cost test items, “My biggest problem is with the open ended questions. I hate them. They are the most difficult. I can get it right, but I can’t explain it, so sometimes I just erase [my work].”

Other students acknowledged that ability beliefs affected their effort. “Sometimes I guess if I don’t understand it. Math is hard. The way they want you to do it is hard. All the answers I had were wrong. It made me feel angry,” and “Sometimes I know I’m not going to get a good grade, so I don’t try.” Of the 11 comments attributed to ability

beliefs, eight were made by students with erratic profiles. Only one student with a low profile attributed lack of effort to ability, "If I don't try it's because they are hard. I would try harder if they were easier."

Table 14. *Attributions for Negative Effort Disaggregated by Assessment Profile and Session*

	High		Low		Erratic		Not Known		
Session	1	2	1	2	1	2	1	2	Total Attribute
Classmates	0	0	0	0	1	0	0	0	1
Lack of focus	1	1	1	3	1	2	0	1	10
Teacher	1	3	0	2	0	3	0	2	11
Ability beliefs	1	0	1	0	6	2	0	1	11
Cost	5	3	5	2	11	2	0	0	28
Attitude	0	0	2	0	1	1	0	0	4
Lack of consequence	0	0	1	0	1	0	0	0	2
Not attributed	1	0	0	1	3	0	0	0	5

Students also often referred to problems with "focus." Sometimes they attributed a lack of focus to classmate distractions. These were coded and reported under the "classmates" category. The remaining focus comments all referred to lack of concentration or an inability to maintain engagement, "[The] problem is more effort. At the end I just start slipping. I lost my focus, especially on the ones I don't know." In some cases, focus was hard to separate from cost, "In math I'm not focused. The math test is

too long. I start doing other stuff." However, when students specified "focus" as the reason for not giving full effort, they were coded under this category.

Teacher qualities were also reasons students cited for failing to give full effort. General comments about teachers were disregarded, leaving concerns about lack of preparation or a failure to meet expressed student need. One student with an erratic pattern said, "I give more effort [in other classes] because the way Mr. Y teaches—you don't understand. He says we're not trying when we are. Sometimes I get frustrated." Another student noticed differences in teacher support, "It's hard to understand. Last year's teacher used to help us one-on-one if you didn't get it. This year we have to do it by ourselves. Mr. Y shows you one time and if you can't get it, that's it." Comments related to lack of preparation were far greater during the second coaching session, increasing from one comment to ten which were fairly evenly divided among all profile groups. One student tried to explain this change, "… [it's] getting harder. In some cases, questions we hadn't learned, in other cases I couldn't remember."

Exit questionnaires. At the conclusion of the intervention, 29 students were administered an exit questionnaire to determine what aspects of the experience they had found most helpful and to measure final attitudes regarding effort and importance. Seven students were absent on the scheduled day. Students were also given the option of responding to four open-ended questions about tests, test effort, and data coaching; sixteen students responded. However, all of the responses on the topics of PSSA, 4Sight, and test effort reiterated information obtained in field notes or were unrelated to the research questions. Therefore, no analysis was conducted on these open-ended questions.

Weighted total scores and percentage of students choosing “strongly agree” or “agree” were computed for each question as reported in Table 15. This table groups questions according to importance, effort, and change. Questions one, two, four, and six on the questionnaire were used to determine student importance. Questions three, five, seven, and eight comprised the effort section. The final three questions on the questionnaire were regarding student change.

Students most strongly agreed with the statement, “Seeing my 4Sight scores makes me care about them.” Students also showed strong agreement to “4Sight test scores are important to me”, “Setting testing goals motivates me to give my best effort on 4Sight tests” and “My 4Sight test scores are a good example of my ability.” The two questions related to parent involvement with 4Sight (questions four and ten on the questionnaire), and the idea that talking to school staff about scores improves motivation (question five) showed the lowest agreement.

On the helpfulness scale, students rated five aspects of data coaching on a six-point scale with six being “very helpful” and one indicating “not at all helpful.” Again, weighted totals were calculated and percent agreement was determined by the adding the percentage of students who weighted the activity as five or six points. Weighted totals were used to rank activities from most to least helpful. Table 16 shows that students reported viewing their assessment profile and setting goals as most helpful, while talking with adults and participating in the large group presentation were ranked as least helpful.

Table 15. *Responses to Exit Questionnaire Regarding Importance, Effort and Change*

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Weighted Total	Percent Agreement
Importance						
4Sight tests are important to me.						
51.7%	34.5%	10.3%	3.4%	0.0%	118	86.2%
Seeing my 4Sight scores makes me care about them.						
51.7%	44.8%	3.4%	0.0%	0.0%	127	96.6%
My 4Sight scores are a good example of my ability.						
34.5%	51.7%	13.8%	0.0%	0.0%	110	86.2%
My parents/guardians tell me that 4Sight test scores are important to them.						
27.6%	17.2%	37.9%	6.9%	10.3%	72	44.8%
Effort						
Setting testing goals motivates me to give my best effort on 4Sight tests.						
24.1%	62.1%	10.3%	3.4%	0.0%	110	86.2%
Talking with teachers and principals about my 4Sight test results motivates me to do better.						
31.0%	34.5%	31.0%	3.4%	0.0%	88	65.5%
I always give my best effort on 4Sight tests						
34.5%	41.4%	17.2%	3.4%	3.4%	103	75.9%
Knowing that teachers use 4Sight tests to see what I already know makes me want to do my best when I take these tests.						
41.4%	37.9%	20.7%	0.0%	0.0%	104	79.3%
Change						
I take 4Sight tests more seriously now than I did last year.						
34.5%	41.4%	20.7%	3.4%	0.0%	101	75.9%
I talk with my parents/guardians about my 4Sight test scores more than I did last year.						
17.2%	27.6%	24.1%	17.2%	13.8%	80	44.8%
I give more effort on 4Sight test this year than I did last year.						
31.0%	44.8%	17.2%	3.4%	3.4%	102	75.9%

Table 16. *Responses to Exit Questionnaire Regarding Helpfulness*

	VH 6	5	4	3	2	NH 1	Weighted Total	Agreement
Seeing my own assessment profile	48%	37.%	6.9%	0.0%	6.9%	0.0%	147	86.2%
Setting goals for the next test	51.%	27.%	13.8%	6.9%	0.0%	0.0%	146	79.3%
Talking about how test data affect my future	62.%	17.%	3.4%	3.4%	3.4%	10.%	137	79.3%
Talking with an adult about my test results	37.%	27.%	20.7%	6.9%	0.0%	6.9%	130	65.5%
Participating in the presentation on testing and effort	13.%	27.%	37.9%	6.9%	10.%	3.4%	108	41.4%

Five students responded to the open-ended item labeled "data coaching." Thirteen comments related to the intervention were also collected in field notes in the final data coaching session and were considered together with comments collected from the questionnaire. Nine students described the action of "seeing" test scores as helpful, while three described "talking" as important in reflecting and goal setting. Three students combined these verbs saying that those actions together were effective. One student said, "It's good to see your scores because it makes you try harder. I'd rather talk to someone so that they can help me [sic]. If I just looked at my scores it wouldn't be helpful. It's

better to talk about it in person." Three students made negative comments about the emotional aspects of failing to meet their own expectations. An example of this feeling was in this comment, "It helps me to see my scores even though they make me cry. It helps me to know how I'm doing."

Results

Research question one. Research question one asked, "What effect does a data coaching intervention have on student test engagement?" Baumert and Demmrich (2001) found that the promise of feedback was not significant in affecting student motivation on low-stakes tests. The results of this study determined that an intervention that included actual feedback was also not significant in increasing student effort. A thorough comparison of pretest and posttest scores and difference between pretest and posttest scores showed lack of significant change for either the experimental or control groups. Had students ascribed greater importance or effort to 4Sight as a result of data coaching, scores would have been expected to increase. Mean scores of the experimental group decreased while the control group stayed relatively stable. The variable of time explained some decrease in all scores, showing that students gave less effort and ascribed less importance over time; in this study, scores were compared over a period of five months. Additional analysis was conducted to determine if significance could be found by comparing students based on assessment profile type. This was done to ensure that ability or ability beliefs were not impacting the findings. Again, all results were found to lack significance. This finding is consistent with a claim by Wise and DeMars (2005) that students with high ability are as likely to report low motivation as students with lower

ability. While students with high achievement profiles reported higher importance at the posttest, their mean effort score showed the greatest decrease.

Research question two. Research question two asked, "What aspects of data coaching do students report as most helpful in increasing task value?" Students reported that they found most aspects of the experience helpful. Students ranked *seeing* their assessment profile as most helpful, with 86.2% of students rating this activity a four or above on a six-point scale. Qualitative data supported this finding. While students also found *talkin*g with an adult helpful, they more frequently found conversations about the impact of their scores on the future and goal setting to be most meaningful. Talking about scores in general was rated as less helpful among questionnaire participants. Students felt that a presentation was the least helpful of all intervention activities.

Research question three. The third research question asked, "In what ways do students report that data coaching affects their attitudes toward low-stakes assessments?" Participants showed agreement or strong agreement with all statements in the exit questionnaire except the two related to parental influence (questions four and ten). Nearly all students (96.6%) agreed or strongly agreed to the statement, "Seeing my 4Sight scores makes me care about them." Overall, 86.2% of participants agreed with the statement, "4Sight tests are important to me," with more than half of all students strongly agreeing. And 86.2% of students recognized that 4Sight tests are an accurate assessment of their ability. Aside from the question about parent discussion, the other two questions in the change category assessed student attitudes new to this year. More than three-fourths of participants revealed that they both take low-stakes tests more seriously, and they gave more effort than they did in previous years.

Validity and Reliability

The SOS (Sundre, 2007) has been administered to over 15,000 students, mostly during low-stakes tests at James Madison University. Several studies have been conducted to assess reliability and validity of this instrument (Thelk, et al., 2009; Wise, 2006) and have generally found the SOS to be valid in measuring student effort and importance on low-stakes tests. The reliability evidence for the total is between .80 and .89 and is consistent even when subsections are used in isolation. Variability of scores is reduced in situations where student orientation is high-stakes. In these situations students are inclined to consider the test important and therefore generally exert more effort

In a pilot study, student coaching sessions were audio recorded and professionally transcribed. Analysis of these data was difficult due to long pauses, nervous laughter, and unintelligible speech. Although questions were scripted, as they were restated to elicit student responses, they became leading—limiting student range of responses and posing a risk to validity. Creswell (2007, p. 208) lists "member checking" as a validation strategy for qualitative data. This technique was employed during data coaching sessions for this study. A standardized form (Appendix D) was used to ensure reliability during the intervention. Notes were taken during interviews, eliminating nervousness effects caused by tape recording. Students reviewed these notes before leaving the session to ensure that the comments recorded were an accurate reflection of their feelings.

A pilot study was also conducted on the exit questionnaire using a sample (n = 20) of ninth grade students who would not be eligible to participate in the research. After students took the exit questionnaire, they met with the researcher as a focus group to give feedback regarding clarity. The questionnaire was effective in eliciting ample data on the

four open-ended questions. During the actual study, this section of the questionnaire netted little new information. This may have been due to student perceptions that the areas of inquiry had been fully discussed during data coaching sessions, whereas the ninth grade students were giving information for the first time.

Summary

Findings regarding the effectiveness of data coaching provided mixed results. SOS results taken during low-stakes tests showed that the intervention was not sufficient in improving aggregate student effort and importance at the experimental site. Mean scores for effort decreased in all cases except in a few cases at the experimental site. There, scores for importance increased for eight students with high or unknown assessment profiles. However, this increase was not significant. Although outliers to the data were found in all pretest and posttest scores at the experimental site (as shown in Figure 5), analysis revealed that they did not impact findings.

Despite the mean decrease in effort, students reported that viewing their test scores, discussing ways that testing impacts them, and setting goals for future test performance were helpful activities. Exit data revealed that most participants agreed that low-stakes tests were important, although fewer agreed that their effort was consistently high. Despite the SOS results, more than 75% of participants ascribed greater importance to low-stakes tests and reported improved effort this year.

Field notes revealed a mix of student reactions and beliefs. When analyzed by achievement profile, high students had mixed feelings about their scores, showing both feelings of pride and remorse for lower-than-expected scores. Nearly all high achieving students ascribed importance to assessments. High students attributed their effort-striving

to a variety of factors, but attributed lack of effort more often to teacher and cost. Lower achieving students were overwhelmingly negative about their performance on tests over the past two years. Those students showed mixed beliefs about test importance, most frequently attributing importance to test preparation. The low-achieving group said that the lack of feedback caused them to feel that low-stakes tests were not important. Those students more often said that they did not give full effort on assessments, citing off-task behavior and cost of completing the work as reasons.

Students with erratic patterns showed nearly every coded feeling with similar frequency. They generally acknowledged test importance, naming feedback and views endorsing testing as integral to learning. This group was far more likely to credit teacher effectiveness with causing them to give effort, and cost with causing reduced effort. Focus, ability beliefs, and teachers were also frequently cited as reducing effort. Overall, more than half the comments for positive effort were attributed to teachers. Qualitative data regarding student feelings about achievement, future goals, and classroom orientation provided some insight.

Chapter 5: Interpretation, Conclusions and Recommended Actionable Solutions

The purpose of this study was to determine the degree that student motivation on low-stakes tests could be influenced, and what activities students reported as most helpful in this endeavor. Three research questions framed this mixed-methods study. Data used to investigate these questions included pretest and posttest SOS scores, field notes, and exit questionnaires. Results of these data were mixed, requiring further interpretation.

Interpretation of Findings and Results

Limitations. One of the limiting factors of this study was the relatively small sample size. The sample generated low power in the findings, impacting generalization and broad application. Working with students at a school site provided the opportunity for rich data, but posed inherent obstacles in attaining a meaningful sample size. Although some parents were wary of allowing their children to participate in research, the greater problem was in student assent. The topic of research may have had a negative effect on student participation. Madaus and Clark (2001) described the impact that voluminous testing has had on high poverty and minority students. Students at those schools have been subjected to less meaningful instructional designs and feel more pressure to perform than their counterparts in other schools. The focus on testing has had the unintended consequence of decreasing student engagement in both learning and testing among other negative effects. Although participation opportunities were discussed with all students and individual consultations were held upon request, some students indicated that they were not interested in discussing their assessment profiles. The need for student assent could not take priority over potential ethical considerations of coercion. Therefore, once students declined, attempts to gain assent were discontinued. Future

research of the effectiveness of coaching would need to include multiple research sites to provide a sample large enough to be significant.

Student attendance was a secondary factor in limiting the sample size. The SOS was designed to be administered during a low-stakes test. Because teachers use 4Sight data to interpret aggregate group achievement, they did not require students to make up tests that were missed. Students who were absent for the November or March 4Sight test did not take the SOS. As a result, five data points were lost, reducing the SOS sample (n = 31). Student interviews, documented in field notes, took place over a series of days and were not impacted by absences, yielding a higher sample (n = 36).

Lastly, the compacted time frame may have limited a positive change in student effort. The plan for this study was to conduct coaching sessions for the longest duration possible. Initial coaching could not commence until the administration of the first 4Sight test, which was set in November by the school district leadership. Some delay occurred in loading tests to the management software for scoring and again in uploading scores to the data warehouse. Additional delays were encountered in IRB approval due to the winter holiday. The timeline was amended to begin coaching at the soonest possible time; however, this did not occur until January. It was impossible to extend coaching sessions further into the school year as administrators understandably wanted students to focus on high-stakes testing which started in mid-March and ran through April. As is likely that changing student beliefs would require more repetition than this study allowed, a future study might test whether the intervention would be more successful over a greater period of time.

Engagement. The first research question considers what effect data coaching has on test engagement, defined here as effort and importance. The most important discovery of this study was the unexpected decrease in reported mean scores for effort and importance in all cases. Although these changes in scores were not significant, there was a significant effect for time $F(1, 60) = 5.831$, $p = .019$ in the scores for effort. In other words, posttest scores tended to be lower on average for all subjects, regardless of treatment group, compared to pretest scores.

Coded field notes gave some insight regarding the difficulty of increasing student motivation on assessments. Expectancy-value theory describes effort as a function of students' expectancies for success (their ability and performance beliefs) and the value that they ascribe to the task. The aim of this intervention was to inform students about the implications of testing in an effort to raise the importance that students associate with testing. Participants gave several reasons for failing to give full effort on assessments. They cited low ability beliefs, negative teacher relationships, and off-task behavior as causes for disengagement. Frequently, however, students described the high cost in terms of time and effort as limiting effort. In fact, 28 of 72 comments (38.8%) attributed cost in time and energy as a reason for failing to give full effort. As students described their test taking behavior, they mentioned that they were more likely to avoid or give minimal attention to "open-ended" test items; those that require written explanations and constructed responses. An example of this behavior was given by a student with a high assessment profile, "I don't like open-ended questions. You have to restate the questions and do a three step solution. You need to give your own opinion and supporting details."

Cost (n = 28) was more often cited than ability (n = 11) in describing low effort in reading. Lower ability students commented on the effort required to read selections while remaining students described the effort required to answer open-ended items. This added to the finding by Sundre (1999) that essay-type items in non-consequential testing conditions caused students to report the lowest motivation, effort, and performance compared to multiple choice items in similar testing conditions, or any format in consequential conditions. In fact, some students (n = 6) in the sample specifically stated lack of consequence as a reason for low motivation, further indicating that the nature of low-stakes testing inherently impacts motivation.

Madaus and Clarke (2001) found lower motivation in minority students who believed that achievement was unattainable or those who viewed tests as irrelevant to their futures. This may explain why students in this sample also indicated that ability beliefs and knowledge impacted their effort on the math sections. Here, students faced test questions that required skills they had not yet acquired, compounding their struggle for full effort. Three students attributed low effort to lack of focus when the underlying problem was lack of skill. An example of this was found in the comment, "[The] problem is more effort. At the end, I just start slipping. I lost my focus: especially on the [questions] I don't know."

Test boredom was another possible explanation for reduced effort over time. Nearly every student recognized the redundancy and over-use of low-stakes assessments. Because 4Sight tests are not true benchmark tests, but instead repeatedly measure students against end-of-year standards, students are often assessed on information they have not yet learned, and questions are repeated over time. One student acknowledged

this in saying, “Sometimes I think they ask the same questions over and over, just with different words. I think that's stupid.” Although students acknowledged on the exit questionnaire that 4Sight tests are important to them, field notes revealed that the cost to complete complex test questions, student ability beliefs, and the repetitive nature of the testing experience caused reduced effort. Coaching students on test importance was not able to compensate for the demands of time and skill required for full engagement.

Increasing task value. The second research question asked, “What aspects of data coaching do students report as most helpful in increasing task value?” The single strongest response (62.1%) was to the statement “Talking about how test data affect my future”; although more students (86.2%) generally agreed that seeing their assessment profile was helpful. Student comments reflected individual preferences regarding what they found helpful. It is clear that some students found discussions about their data helpful, while others would have preferred to view their results privately. Data on student feelings may explain this difference.

As reported in Table 10, students with high profiles shared equally in positive and negative feelings about their results. Four students said that they were proud while two students each said that they felt smart or successful. One comment was attributed to a general feeling of “happy”. All the negative comments (n = 9) were attributed to worry, regret, anger, or surprise. In every case, these students either liked seeing their results or discussed that they wanted to improve. In contrast, of the 23 statements regarding feelings for students with a low assessment profile, none was positive. Students reported disappointment (n = 5), regret (n = 3), frustration (n = 2), embarrassment or shame (n = 2), anger (n = 1) and worry (n = 1). Five students said that their scores made them feel

"bad" and four students cried. Students with an erratic profile had a wider range of reactions. One or two comments were counted for all 10 positive feelings and one or two negative comments were made for 10 negative feelings. None of the erratic students described the negative feelings of disappointment and embarrassment

Viewing assessments was difficult for the low assessment profile group and for some of the erratic group. This may explain why some students preferred "seeing" their data while others wanted to "talk" about their profiles. It was understandable that students with low achievement might want to avoid the negative feelings they have when confronting low test scores. However, if students are indeed giving full effort and are not able to achieve their goals, these negative feelings may add an emotional cost to the assessments that could decrease their effort to strive on future tests. School personnel must consider the negative emotions that lower achieving students may have when discussing test scores. They should emphasize growth rather than achievement and focus on achievable mastery goals.

Perhaps predictably, students rated a presentation on testing and test effort as least helpful. This presentation lasted approximately 60 minutes and was implemented to establish common language for students. Although designed as in interactive discussion, some students were disengaged, particularly in the seventh grade group. This lack of engagement resulted in some minor disruption to the flow of discussion. The eighth grade group was much more interactive, although one student asked to be excused, claiming boredom. Subsequent data coaching sessions revealed that students had to be reminded of the content from the presentation, so the benefit may have been minimal. Future studies should consider a more effective way to educate students on the topics of test culture and

effort. This may include small group discussions or teacher directed mini-lessons perhaps as part of a morning meeting in homerooms.

Student Attitudes. The final research question used exit questionnaire data to consider how data coaching affects student attitudes on low-stakes assessments. Students overwhelmingly (96.6%) claimed that seeing their scores made them care about them. This was supported qualitatively by seven students who said that they felt that low-stakes tests were unimportant in the past because they had not received any feedback. Students also reported that they took low-stakes tests more seriously this year than they did last year (75.9%), and that they gave more effort this year compared to last (75.9%). Data coaching appeared to have little value in involving parents in low-stakes assessments as fewer than half (44.8%) of students reported talking with their parents about these tests more this year. This was not surprising given that parents did not participate in data coaching apart from signing permission forms, and parents were rarely given as a reason that students ascribed importance to low-stakes tests (n = 2).

Initially it appeared that the data from the exit questionnaires contradicted the findings from changes in SOS scores at the experimental site. These results indicated that student effort and importance decreased over the year. But the exit questionnaire asked students to consider changes in their beliefs from the prior year, not from the beginning of the year. While there may be some impact from students answering questions to reflect what they perceive as a desirable response, qualitative data showed that most students were generally positive about the experience. Data also indicated that student attitudes were complex and were influenced by several factors.

***The Effect of Consequence*.** Given the mean reduction in scores for the experimental group it was interesting to note that when importance scores were disaggregated, students in the high and not known assessment profile categories were alone in showing an increase in importance. The high group (n = 6) increased in pretest to posttest scores from 19.5 (SD = 5.167) to 20 (SD = 3.286) and the not know group (n = 2) increased from 17 (SD 5.657) to 19 (SD = 2.828). While the increase in importance for the high group was accompanied by the largest decrease in effort, from 20.5 (SD = 1.517) to 15.33 (SD = 5.715), it turned out that the decline in mean effort for this group was largely due to a single seventh grade score. In such a small sample (n = 6), data was more sensitive to skew from outliers. Neither the change in importance $t(5) = -.311$, $p = .768$, nor the change in effort $t(5) = 2.289$, $p = .071$ for the high group turned out to be significant. With only two data points, the not known profiles cannot be considered significant either. Still, the result called for investigation to determine why so few students increased in importance while the mean scores decreased.

Of the eight students identified in these two profile groups, seven were in eighth grade. The current practice in this school district was for transitioning eighth graders to apply for placement at various high schools. These included a vocational school, an art magnet school, and a science and technology high school. In all cases, placements were competitive. During student interviews, many eighth grade students, and all seven of the "high" students, indicated that they were involved in the high school application processes. All applicants noted that assessment data would be used as part of the selection process. Students also indicated value in 4Sight as preparation for high stakes assessments. Of the eleven importance attributions made by the "high" group, eight

(72.2%) were designated as preparation for PSSA or consequence in high school selection.

Expectancy-value theory would predict that students who chose to compete for high school placements would ascribe greater importance to the measures that would determine their success. Supporting this theory, O'Neil, et al.(2005) found that high school seniors preferred certificates of merit for high test scores more than monetary rewards, indicating that certificates would be more valuable in the college application process. This finding may explain why students who were applying to high schools and were dependent on higher test performance reported greater value in 4Sight assessments as compared with all other students.

Further analysis of SOS results by assessment profile revealed and interesting trend. For the experimental population (n = 31), the range of posttest importance scores covered all possible scores: a 25-point score indicating the highest value and a 5-point score indicating the lowest possible value. Table 17 shows descriptive statistics for posttest importance scores for the experimental group by assessment profile. This data indicated that the high profile students had greater mean importance scores that were more evenly distributed, reflecting the relative utility value for those students, and showing the greater range in utility among the other profiles.

Table 17. *Summary Statistics for Posttest Importance scores by Assessment Profile*

	n	Minimum	Maximum	M	SD
High	6	16	25	20.000	3.286
Low	11	6	25	18.091	5.127
Erratic	12	5	22	17.500	5.351
Not known	2	17	21	19.000	2.828

Coded statements supported the idea that students with lower profiles ascribed less importance to 4Sight tests than students with high or erratic profiles. Of the 23 importance statements made by 11 "low" students, more than half (52.2%) were coded as negative. Most of these comments were attributed to lack of feedback (n = 5). Others included lack of consequence (n = 3), general statements (n = 2), and teacher influence (n = 1). In contrast, of 12 statements made by "high" profile students, nearly all (91.7%) indicated value attributed to consequence (n = 4), preparation for PSSA (n = 4), and testing as an integral part of learning (n = 2). One comment gave no attribution. The variety of attributions and the range of scores among the profiles demonstrated that for this sample, test value and "stakes," was highly individualized, even for students within a school.

The Effect of Classroom Orientation on Effort. Although importance scores varied according profile types, student effort scores decreased in all instances and to a greater degree with the sample at School A. It is unlikely that the cause for the decrease is attributable to the intervention, as decreases were found in the control. Additionally, students reported in exit questionnaires that they generally found the data coaching experience to be helpful. The changes in effort scores were not significant, most likely due to the small sample size. However, the effect of time was found to be significant in explaining the change in scores $F(1,60) = 5.831$, $p = .019$. Analyzing effort scores by student profile did not add to the explanation, since Table 9 showed that profile group differences were not found to be significant. Despite the lack of significance, it was concerning that student effort declined over a period of four months, and that the change

in effort scores for School A (M = -1.742, SD = 5.413) was higher than at School B (M = -0.710, SD = 4.002).

A more likely explanation for the differences in these changes was the presence of a confounding variable. One of the limiting aspects of educational research is that it is difficult to control for all factors, as each classroom is comprised of people who espouse a range of behaviors and attitudes that impact their thinking and decisions. While test fatigue may have accounted for the decrease in all effort scores, it did not explain the larger decrease that occurred at School A.

Field notes regarding attributions for effort offer some insight. Students who reported positive test effort gave several reasons. In addition to ability beliefs that increased their expectancy for success, students named parents and peers as influencing their perceptions of test importance. Most often, however, students attributed positive test effort to their teacher. Comments for effort were attributed to teachers when students cited specific beliefs or teaching strategies that reduced the cost of completing the task.

Other statements related to teachers were not coded specifically for test effort as they described more general helping behaviors, but the position of those comments in conjunction with discussion of test effort indicated that students saw a connection between effective teaching and achievement. Although most comments were made regarding teachers of classes that participated in 4Sight testing, some students spoke more generally and even credited their successes to instruction by teachers in prior years. Beyond merely liking them, students were able to articulate specific attributes that supported learning and testing. Most often students described situations where teachers provided extra support to individuals, accepted shared responsibility for learning, and

showed personal interest in student success. One student relayed this account, “My teachers told me to start asking questions. Just to be a smart aleck I started asking questions. But it became a habit and now I understand more.” Another student attributed teacher care to achievement, “Now we have teachers who are focused on our future. They’re looking out for us.” To the extent that teachers influence ability beliefs, they may also impact test effort.

Although cost was most often cited as a reason for failing to give full effort, negative effort was attributed to teachers in 11 comments. In those comments, students described adversarial relationships as negatively affecting effort in class and during tests. Examples of teacher behavior that students reported as having a negative influence on their effort or achievement included: refusing to remediate, sacrificing instructional time to attend to non-instructional tasks, blaming students for lack of achievement, and using learning tasks as punishments. Surprisingly, students claimed to like teachers even when they did not feel that they were effective.

Students reported that lack of instruction and failure to remediate required increased effort during testing or affected student ability beliefs. This was particularly true in math where the nature of the test required students to use information or skills they had not yet acquired. Three students claimed this was due to a lack of instruction when the problem may have been better understood as a function of the test construction. In any event, testing students on information they had not learned affected student ability beliefs, and performance beliefs and had particular impact on a process-oriented subject like math. One student described the frustration he felt during a math test:

> I guessed on some of the questions in math. I eliminated the questions to two choices, then [I] did eeny, meeny, miney mo. I just didn't have the knowledge. Some of the questions we hadn't learned. They felt like eighth grade questions. Three weeks later we learned some of the stuff and it was so easy. I felt stupid for having gotten it wrong.

Positive and negative attributes for teacher influence indicated increases between coaching sessions. This may have been due to the fact that students were reflecting on a recent test in the second coaching session, and that they were preparing for PSSA testing. When teachers were named, all positive comments were attributable to a single teacher (Teacher X) and all negative comments were attributable to a different teacher (Teacher Y). Teacher names have been excluded and are hereafter referred to as Teacher X and Teacher Y. Although there were four core teachers on the team, those two teachers were the ones teaching courses assessed by 4Sight. At times, students commented positively about other teachers in general, but all the negative remarks were directed at Teacher Y. Two negative effort comments regarding Teacher Y compared his instructional style with that of a former teacher who had been viewed as a highly effective instructor who would "explain things in several different ways." Students were surprisingly objective about their teachers, even when being critical, focusing on the teaching rather than the teacher:

> When teachers don't make us part of the lesson it's less motivating. Teaching style doesn't really matter in motivation. The problem is more that Teacher Y doesn't teach in a way we can understand. If we ask a question, he doesn't really answer it but goes back and says the whole thing over again.

The fact that Teacher Y was new to the staff and did not have a preexisting reputation at the beginning of the year may explain why negative comments increased as time went on.

Goal structures are defined by the way teachers and students approach learning. Although researchers have defined goal structures in a variety of dichotomies, most are similar enough to be described as "mastery" or "performance" (Pintrich & Schunk, 2002). Mastery goal orientation is marked by a focus on developing new skills and personal achievement. In contrast, performance orientation is competitive—students work for favorable judgments in comparison to others. Mastery goal orientation is desirable particularly among low achieving students as it is more likely to have a positive effect on student achievement (Meece, et al., 2006).

Although classroom observations were not conducted, student reports of classroom activities and descriptions of teacher behaviors indicated that Teachers X and Y have different goal orientations. Students reported that they believed Teacher X wanted them to succeed and was willing to work with them to improve. Students described strategy instruction, but also instances when Teacher X remediated instruction until they "got it." By contrast, students felt judgment from Teacher Y. They were not comfortable asking for help, nor felt that the help they got was productive:

> I think the problem with my grades is my ability. My grades are not good. I stay after school for [help] but we don't really get help unless I ask. I give more effort in Mr. X's class [than in Mr. Y's class] because of the way Mr. Y teaches; [I] don't understand. He says we're not trying even when we are.

Martin and Dowson (2009) found that student motivation increased in classrooms with a supportive learning climate. Comments from students in this sample supported this claim.

Expectancy-value theory predicts that students are more likely to give effort when they believe they will do well and value the task. Teachers have influence in both these areas. Students are more likely to adopt mastery goal orientation when they see their teachers emphasizing understanding over competition. Classrooms that value competition for grades and compare students' abilities are more likely to foster performance goals (Anderman & Midgley, 1997). Teachers who want to maximize student effort should employ practices that support individual mastery and reduce emphasis on competition and performance.

Conclusion

Expectancy-value theory describes student effort as a construct of both expectancies for success (ability and performance beliefs) and task value (importance, usefulness, intrinsic and utility value, and cost). Studies have indicated that the transition to middle school brings new challenges (Anderman & Midgley, 1997; Eccles, et al., 1983; Jacobs, et al., 2002). At a time when student ability beliefs are waning, classroom goal structures are also turning from mastery orientation to performance orientation in many cases (Anderman & Midgley, 1997). The purpose of this study was to determine whether data coaching could improve task value on low-stakes measures of achievement.

Question one. Question one inquired about the effect of data coaching on student test engagement. SOS scores indicated that the intervention was not effective in improving effort and task importance. Although extensive tests were run to find significance, in all cases the null hypothesis was retained.

Time explained change in some scores in that effort decreased in both the control and experimental samples. Disaggregated data at the experimental site revealed that high achieving students, and particularly eighth graders who were applying to competitive high schools, reported greater importance in the post measure. Discussions with students about the application process revealed that although 4Sight tests have been characterized as low-stakes assessments, some students ascribe greater importance as they recognize potential usefulness in performance. This is an example of how test consequence may vary among students even within a single classroom. It could also be argued that for some students pleasing parents or teachers could be considered consequential if failure to do so would bring negative feelings or results. It should be recognized that determining the "stakes" of an assessment is relative to the personal value for each student. This may explain, in part, why increasing aggregate importance is difficult.

Question two. The second question considered which elements of the data coaching experience student find most helpful. Table 16 showed that students had strong agreement with statements of helpfulness for viewing their data, setting goals, and talking about how data affects their futures. Talking with an adult about test scores had a weaker reaction, while viewing a presentation was endorsed as least helpful.

This finding seemed to conflict with the findings of question one. Although more than 75% of students found viewing their assessment profiles, setting goals for future tests, and talking about how test data affects their future as helpful, it did not appear that these activities influenced motivation. This discrepancy may be explained by data regarding student feelings. For some students, particularly those with low assessment profiles, reviewing disappointing test results was difficult. Other students, however, felt

proud or successful. This range of feelings may have explained why some students preferred merely viewing their data while others were eager to talk about their profiles.

Question three. The final research question asked, "In what ways do students report that data coaching affects their attitudes toward low-stakes assessments?" Table 15 showed that 75% of students noted a change in importance and effort from last year to this year. Those students agreed that they take low-stakes tests more seriously and that they gave more effort this year. While mean effort and importance scores declined over the course of the intervention this year, effort and importance were not assessed last year.

Analyses of field notes explained why students felt tests were or were not important, and why they gave or failed to give full effort. While students reported varying factors that influenced effort, cost of task completion and teacher influence were most often cited. In both coaching sessions, students named positive teacher actions most often as increasing effort. Although cost was the most frequently discussed negative reason in the first coaching session, teacher action was named more often in the second. Negative teacher influence may have been a confounding variable in the decrease of student effort at the posttest. This was explained through student comments that revealed a performance-oriented approach in one classroom.

Data coaching was implemented as an outside intervention in this study. It was possible that the benefits students derived in setting goals and reflecting on their performance was countered by a performance orientation in the classroom. Future research should consider whether this intervention is effective when implemented by teachers in a classroom setting as part of a mastery goal orientation. This strategy could

also provide the opportunity for coaching students during the full school year (or even across years), eliminating the effects from time limitations.

Recommendations

Several factors limited the scope of this study and should be considered in future research. The first limit was in the sample size. With only 36 participants, data was limited. This was particularly true in SOS results and was compounded by student absences, further limiting the data set to 31. This impacted power and therefore significance. The second limiting factor was time. The parallel demands of research studies and school assessment calendars reduced the duration of the intervention, potentially reducing significant results. Future studies should determine whether widespread implementation of a data coaching over a greater length of time, and with a broader sample, could increase significance. A longitudinal study could determine if data coaching is effective in mitigating the decline in motivation as students enter adolescence.

This study confirms prior studies that showed the relative difficulty of improving test effort among students who lack personal consequence for their performance (Sundre, 1999; Wise & DeMars, 2005). Brown and Walberg(1993) noted that school culture may have been a confounding variable when they studied the motivational effects of test directions. Future research should determine what extent school culture, defined here in terms of classroom goal orientation, has in improving test effort and importance on non-consequential testing.

The findings of this study are instructive to school personnel who wish to maximize student effort in testing situations. Using the Student Opinion Scale during test

administration was an unobtrusive way to easily measure student effort and importance. As expectancy-value theory explains, effort is dependent upon student expectancies for success and their perceived value. In all test situations, maximum student effort is required for test validity. Using the SOS to gauge trends in motivation would enhance instructional practices as well as assist in valid interpretation of test scores, particularly in non-consequential test situations. When allowable, filtering out scores of unmotivated students would provide teachers with better information on which to base teaching decisions (Wise & DeMars, 2005).

As students most frequently cited cost as the reason for their failure to give full effort, schools should carefully consider what assessments are necessary to give and ensure that curricula are aligned to assessment material. In this research study, testing students on unfamiliar information not only impacted their ability beliefs and caused disengagement from future tests—it also gave teachers meaningless information and wasted instructional time. Because most current testing practices are mandated, schools should carefully consider the frequency and use of non-mandated tests, particularly those that are low-stakes.

Lastly, qualitative data endorsed the use of mastery goals in promoting full student effort. The nature of low-stakes tests inherently reduced student task value. One way to mitigate this reduction is by ensuring that the cost of completing assessments is as low as possible. Effective, supportive instruction minimizes energy costs and improves ability beliefs. Increasing task value and expectancies for success improves the likelihood that students will fully engage. As studies have found that effort impacts test performance (O'Neil, et al., 2005; Sundre, 1999; Wise & DeMars, 2005), teachers and school

administrators interested in improving achievement results will find that ensuring mastery orientation in classrooms is an effort factor that they can influence.

Summary

Testing students as a means of determining educational productivity is a standard educational practice that has become more significant since the establishment of NCLB. Increasingly, assessments have become accountability measures for schools; a trend that will become even more personal to teachers in Pennsylvania as the state embarks on a new teacher evaluation tool, which uses student achievement data as a "significant factor" in performance ratings (Pennsylvania Department of Education, 2012). Previous studies have documented threats to score validity when students do not give full effort (Wise, 2009), and that low-stakes tests, in particular, underestimate student knowledge (Baumert & Demmrich, 2001; Duckworth, et al., 2011; O'Neil, et al., 2005; Wise & DeMars, 2005).

This study adds to previous research that investigates the extent that student motivation can be improved as a means of increasing test score validity. Quantitative data indicated the relative difficulty in affecting change through feedback and goal setting alone. Educators should be encouraged, however, by student response to teacher behaviors that are viewed as supportive and committed to mastery. These measures, carried out by those closest to students in the context of a classroom setting, may offer the best solution to the effort problem by improving student task value and expectancies for success.

List of References

1. American Educational Research Association. (2000). Position statement of the American Education Research Association concerning high-stakes testing in preK-12 education. *Educational Researcher, 29*(8), 24-25.

2. Amrein-Beardsley, A. (2009). The unintended, pernicious consequences of "staying the course" on the United States' No Child Left Behind policy *International Journal of Education Policy & Leadership, 4*(1-11), 1-13.

3. Anderman, E. M., Anderman, L. H., & Griesinger, T. (1999a). The relation of present and possible academic selves during early adolescence to grade point. *Elementary School Journal, 100*(1), 3.

4. Anderman, E. M., Maehr, M. L., & Midgley, C. (1999b). Declining motivation after the transition to middle school: schools can make a difference. *Journal of Research & Development in Education, 32*(3), 131.

5. Anderman, E. M., & Midgley, C. (1997). Changes in achievement goal orientations, perceived academic competence, and grades across the transition to middle-level schools. *Contemporary Educational Psychology, 22*(3), 269-298. doi: 10.1006/ceps.1996.0926

6. Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W.H. Freeman.

7. Bangert-Drowns, R. L., & Kulik, C. C. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213.

8. Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research, 53*(4), 571-585.

9. Bangert - Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213.

10. Baumert, J., & Demmrich, A. (2001). Test motification in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(441-462).

11. Braun, H., Chapman, L., & Vezzu, S. (2010). The black-white achievement gap revisited. *Education Policy Analysis Archives, 18*(21), 1-95.

12. Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research, 86*(3), 133.

13. Brunner, M., Artelt, C., Krauss, S., & Baumert, J. (2007). Coaching for the PISA test. *Learning & Instruction, 17*(2), 111-122. doi: 10.1016/j.learninstruc.2007.01.002

14. Cochran-Smith, M., & Lytle, S. (2006). Troubling images of teaching in No Child Left Behind. *Harvard Educational Review, 73*(4), 668-668-697.

15. Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage Publications.

16. Creswell, J. W. (2008). *Educational research: planning, conducting, and evaluation quantitative and qualitative research*. Upper Saddle River, NJ: Pearson.

17. Crocco, M. S., & Costigan, A. T. (2007). The narrowing of curriculum and pedagogy in the age of accountability. *Urban Education, 42*(6), 512-535.

18. Darling-Hammond, L. (2009). Recognizing and enhancing teacher effectiveness. *International Journal of Educational & Psychological Assessment, 3*, 1-24.

19. Darling-Hammond, L. (2010). *The flat world of education: How America's comitment to equity will determine our future*. New York, NY: Teachers College Press.

20. Davies, R. S. (2008). AYP accountability policy and assessment theory conflicts. *Mid-Western Educational Researcher, 21*(4), 2-8.

21. Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18*(1), 105-115.

22. Deci, E. L., & Ryan, R. M. (1996). Need satisfaction and the self-regulation of learning. *Learning & Individual Differences, 8*(3), 165.

23. Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist, 26*(3/4), 325.

24. Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 108*(19), 7716-7720.

25. Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Achievement and achievement motivation. In J.T. Spence (Ed.), *Expectancies, values and academic behaviors* San Francisco, CA: W. H. Freeman.

26. Eckes, S. E., & Swando, J. (2009). Special Education Subgroups Under NCLB: Issues to Consider. *Teachers College Record, 111*(11), 2479-2504.

27. Harter, S. (1981). A new self-report scale of intrinsic versus extrinsic orientation in the classroom: Motivational and informational components. *Developmental Psychology, 17*, 300-312.

28. Harter, S., Whitesell, N., & Kowalski, P. (1992). Individual differences in the effects of educational transitions on young adolescent's perceptions of competence and motivational orientation. *American Educational Research Journal, 29*(4), 777-807.

29. Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. doi: 10.3102/003465430298487

30. Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development, 73*(2), 509-527.

31. Johnson, C., & Kritsonis, W. A. (2010). The achievement gap in mathematics: A significant problem for African American students. *Doctoral Forum, 7*(1), 1-12.

32. Lepper, M. R., Iyengar, S. S., & Corpus, J. H. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of Educational Psychology, 97*(2), 184-196. doi: 10.1037/0022-0663.97.2.184

33. Madaus, F., & Clark, M. (2001). The adverse impact of high stakes testing on minority students: Evidence from 100 years of test data. In G. Orfield & M. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high stakes testing in public education.* New York: The Century Foundation.

34. Maier, S. R., Seligman, M. E., & Solomon, R. L. (1968). Pavlovian fear conditioning and learned helplessness: Effects of escape and avoidance behavior of (a)the CS-US contingency and (b) the independence of US and instrumental responding. In B. A Campbell & R.M. Church (Ed.), *Punishment and aversive behavior*. New York: Appleton-Century-Crofts.

35. Marcoulides, G. A., Gottfried, A. E., Gottfried, A. W., & Oliver, P. H. (2008). A latent transition analysis of academic intrinsic motivation from childhood through adolescence. *Educational Research & Evaluation, 14*(5), 411-427. doi: 10.1080/13803610802337665

36. Markus, H., & Nurius, P. (1986). Possible selves. *American Psychologist, 41*(9), 954-969.

37. Martin, A. J., & Dowson, M. (2009). Interpersonal relationships, motivation, engagement, and achievement: Yields for theory, current Issues, and educational practice. *Review of Educational Research, 79*(1), 327-365.

38. Meece, J. L., Anderman, E. M., & Anderman, L. H. (2006). Classroom goal structure, student motivation and academic achievement (Vol. 57, pp. 487-503).

39. Miechenbaum, D. (1985). Teaching thinking: A cognitive-behavioral perspective. In S. F. Chipman, J. W. Segal & R. Glaser (Eds.), *Thinking and learning skills, Vol 2: Research and open questions.* Hillsdale, NJ: Lawrence Earlbaum Associates.

40. Murnane, R. J. (2007). Improving the education of children living in poverty. *Future of Children, 17*(2), 161-182.

41. O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*(3), 185-208. doi: 10.1207/s15326977ea1003_3

42. O'Neil, H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3*(2), 135.

43. Ogbu, J. U., & Simons, H. D. (1998). Voluntary and involuntary minorities: A cultural-ecological theory of school performance with. *Anthropology & Education Quarterly*, *29*(2), 155.

44. Pennsylvania Department of Education. (2012). Educator effectiveness project. Retrieved May 15, 2012, from http://www.portal.state.pa.us/portal/server.pt/community/newsroom/7234/teacher_evaluation/1034646

45. Pennsylvania Training and Technical Assistance Network. (2011). 4Sight Benchmarks Retrieved February 12, 2011, from http://www.pattan.net/teachlead/AssessingtoLearn.aspx

46. Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*(4), 667-686. doi: 10.1037/0022-0663.95.4.667

47. Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research and applications* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

48. Ramirez, A., & Carpenter, D. (2005). Challenging assumptions about the achievement gap. *Phi Delta Kappan, 86*(8), 599-603.

49. Steinberg, L., Dornbusch, S., & Brown, B. (1992). Ethnic differences in adolescent achievement. *American Psychologist, 47*(6), 723.

50. Sundre, D. L. (1999). *Does Examinee Motivation Moderate the Relationship between Test Consequences and Test Performance?*

51. Sundre, D. L. (2007). *The student opinon scale: A measure of examinee motivation*. Harrisonburg, VA: The Center for Assessment and Research Studies.

52. Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14*(1), 8-9.

53. Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. *JGE: The Journal of General Education, 58*(3), 129-151.

54. United States Department of Agriculture. (2009). *Income eligibility guidelines*. Washington, D.C.: Federal Register Retrieved from http://www.fns.usda.gov/cnd/Governance/notices/iegs/IEGs09-10.pdf.

55. United States Department of Education. (2008). *Digest of Education Statistics: 2009*. Washington, D.C.: Retrieved from http://nces.ed.gov/programs/digest/d09/tables/dt09_072.asp?referrer=lis.

56. United States Department of Education. (2010). The condition of education 2010. Washington, DC: United States Department of Education.

57. Unrau, N., & Schlackman, J. (2006). Motivation and its relationship with reading achievement in an urban middle school. *Journal of Educational Research, 100*(2), 81-101.

58. Warner, C. B., & Phelps, R. E. (2008). The relationship between motivational orientation and educational aspirations in urban, African American youth. *Middle Grades Research Journal, 3*(2), 71-85.

59. Weiner, B. (1972). Attribution theory, achievement motivation, and the educational process. *Review of Educational Research, 42*(2), 203-215.

60. Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*, 548-573.

61. White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review, 66*(5), 297-333.

62. Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review, 30*(1), 1-35. doi: 10.1016/j.dr.2009.12.001

63. Wigfield, A., & Eccles, J. S. (2000). Expectancy--value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68-81. doi: 10.1006/ceps.1999.1015

64. Wigfield, A., Eccles, J. S., Mac Iver, D., Reuman, D. A., & Midgley, C. (1991). Transitions during early adolescence: Changes in children's domain-specific self-perceptions and. *Developmental Psychology, 27*(4), 552.

65. Wigfield, A., & Wentzel, K. R. (2007). Introduction to motivation at school: Interventions that work. *Educational Psychologist, 42*(4), 191-196. doi: 10.1080/00461520701621038

66. Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*(2), 95-114. doi: 10.1207/s15324818ame1902_2

67. Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *JGE: The Journal of General Education, 58*(3), 152-166.

68. Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17. doi: 10.1207/s15326977ea1001_1

69. Wise, S. L., & Xiaojing, K. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183. doi: 10.1207/s15324818ame1802_2

Appendix A

The Student Opinion Scale (Sundre, 2007)

This survey is part of a doctoral study designed to examine test effort.
The purpose of this instrument is to determine student effort and importance on low stakes assessments.
Please complete this information before completing the scale:
A. What day of the month is your birthday? (This should be a number between 1 and 31.)
__
B. Who is your homeroom teacher? _____
C. What is the first letter of your middle name? ___
D. What is the first letter of your mother's name? ___

Please think about the test that you just completed. Mark the answer that best represents how you feel about each of the statements below by circling it.

1. Doing well on this test was important to me.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

2. I gave good effort throughout this test.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

3. I am not curious about how I did on this test compared to others.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

4. I am not concerned about the score I receive on this test.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

5. This was an important test to me.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

6. I gave my best effort on this test.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

7. While taking this test, I could have worked harder on it.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

8. I would like to know how well I did on this test.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

9. I did not give this test my full attention while completing it.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

10. While taking this test, I was able to keep working on each question until it was finished.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Student Opinion Scale Administration Directions

You are administering a form of the Student Opinion Survey (Sundre, 2007). This assessment will give an aggregate measure of student engagement and effort on a low-stakes test; in this case, 4Sight. Students must take this survey immediately following the test session. Since 4Sight is not timed, if all students are not finished prior to the end of the class period, you should stop the assessment 5 minutes prior to the end and administer this survey to all students.

First, ask students to answer the 4 questions at the top of the survey.

Say, "You are about to take a survey. It is important that you answer the questions as honestly as you can. This survey is anonymous, and will not be graded. No one will know how you answered the questions and I will not look at the answer sheets. Please answer questions A, B, C, and D at the top of the page first. This letter and number combination will take the place of your name for the researcher. Raise your hand if you need help with these 4 questions."

Once everyone has completed the four questions, say, *"Please think about the test or tests that you just completed. Mark the answer that best represents how you feel about statements 1 through 10 below by circling it. You have five minutes to complete this instrument."*

If students do not understand the wording of a question, tell them to interpret it the best that they can and remind them that the survey is not graded.

Collect all surveys and place them in the accompanying folder.

Appendix B

Exit Questionnaire

Reflecting on the data coaching experience, respond to each statement:

1. 4 Sight test scores are important to me.				
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

2. Seeing my 4 Sight test scores makes me care about them.				
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

3. Setting testing goals motivates me to give my best effort on 4 Sight tests.				
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

4. My parents/guardians tell me that 4 Sight test scores are important to them.				
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

5. My 4 Sight test scores are a good example of my ability.				
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

6. Talking with teachers and principals about my 4 Sight test results motivates me to do better.				
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

7. I take 4 Sight tests more seriously now than I did last year.				
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

8. I always give my best effort on 4 Sight tests.				
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

9. Knowing that teachers use 4 Sight tests to see what I already know makes me want to do my best when I take these tests.				
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

10. I give more effort on 4 Sight tests this year than I did last year.				
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

11. I talk with my parents/guardians about my 4 Sight tests scores more than I did last year.				
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

In this section you will rate activities using a helpfulness scale.
Example: When studying for a test, rate the following environmental factors

	Very helpful					Not at all helpful
Sitting at a desk						
Using a highlighter						

Data coaching is a term used to describe the lessons and individual goal setting sessions you have participated in. These activities were intended to teach you about testing. Rank the following aspects of data coaching to show how helpful you found the experience. This information will help us determine which activities to continue and which to eliminate.

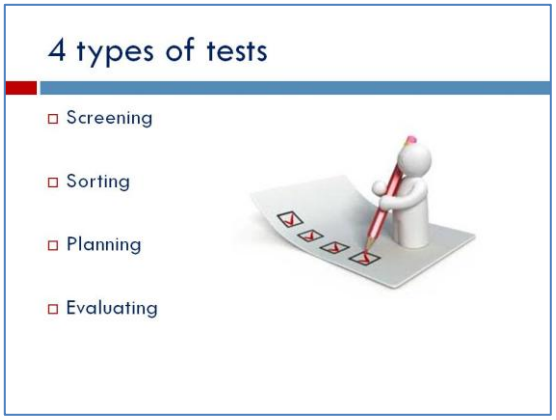
	Very helpful					Not at all helpful
Learning about the 4 levels of proficiency						
Seeing my own Performance Tracker profile						
Talking with an adult about my test results						
Understanding how test data affects my future						
Graphing my goals						

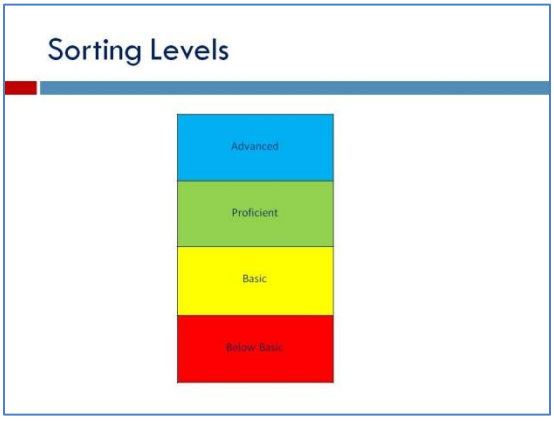
Please share your opinions on the following topics: (optional)

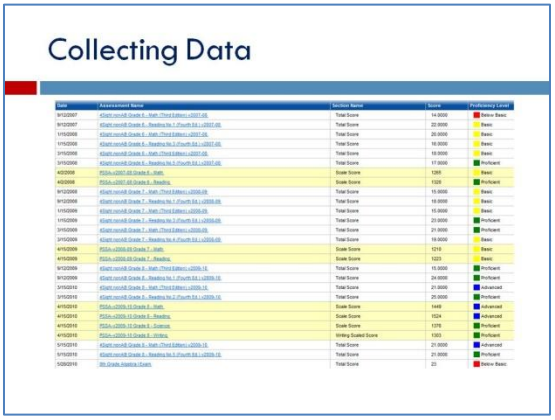
4 Sight tests:
PSSA
Data coaching
Test effort

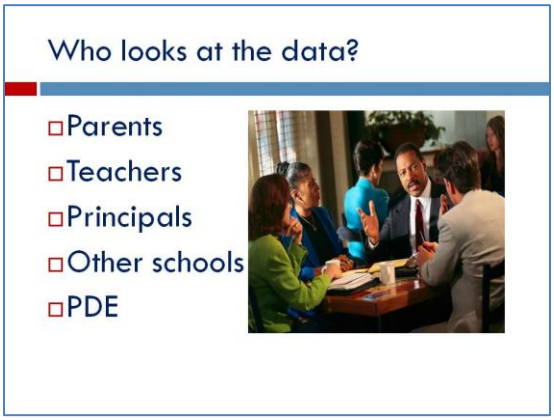
Appendix C

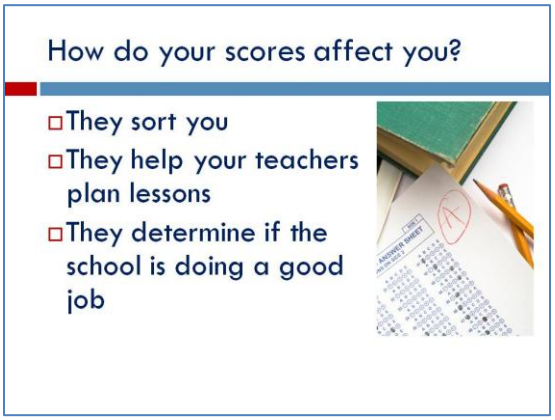
Whole Group Presentation Slides



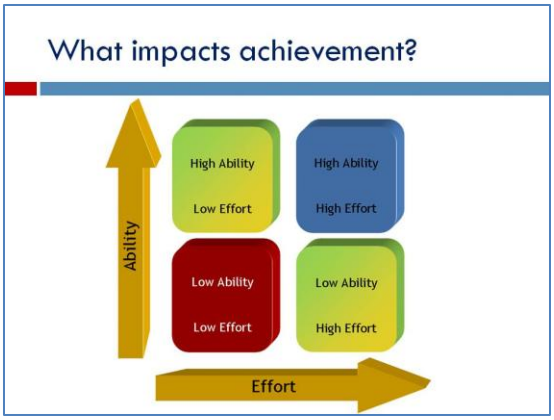


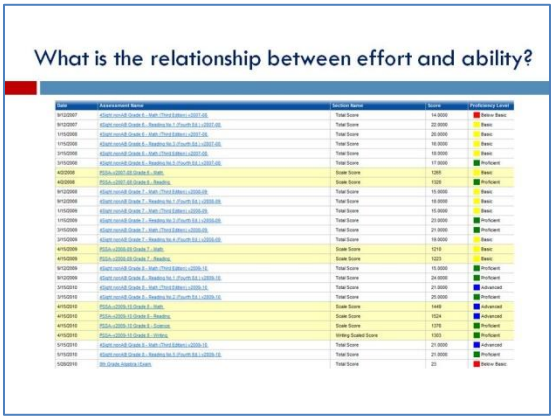


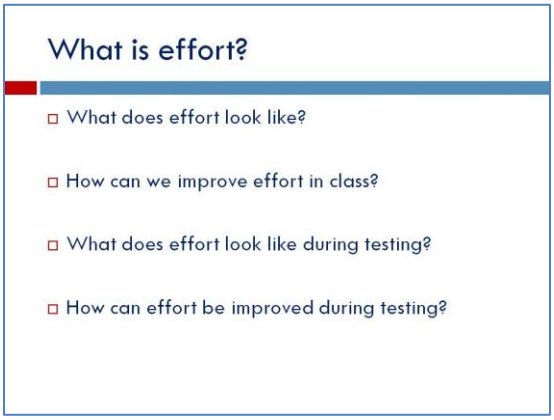


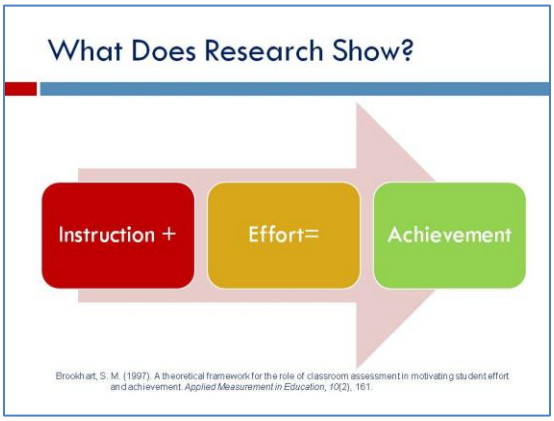


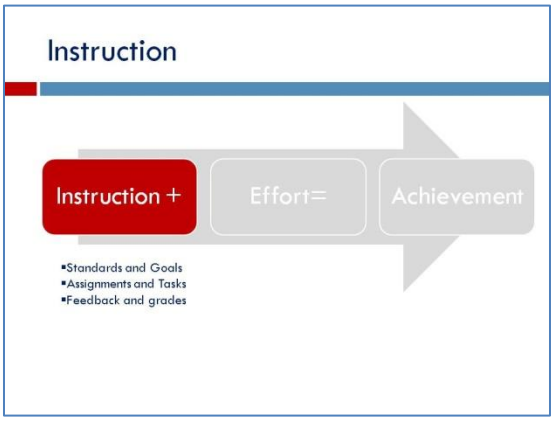

What is Effort?


What impacts achievement?
High Ability
Low Effort
High Ability
High Effort
Low Ability
Low Effort
Low Ability
High Effort
Ability
Effort


What is the relationship between effort and ability?


What is effort?
What does effort look like?
How can we improve effort in class?
What does effort look like during testing?
How can effort be improved during testing?


What Does Research Show?
Instruction +
Effort=
Achievement
Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. Applied Measurement in Education, 10(2), 161.


Instruction
Instruction +
Effort=
Achievement
•Standards and Goals
•Assignments and Tasks
•Feedback and grades

Effort

Instruction + | Effort= | Achievement

- Is based on how much importance and value students give to the assignment or test
- Depends on the student's **self efficacy**, or beliefs about their own abilities
- Is defined by mental effort and effort of actions (Mindfulness)

Achievement

Instruction + | Effort= | Achievement

- Knowledge
- Thinking
- Skills
- Attitudes

Self-efficacy in middle school

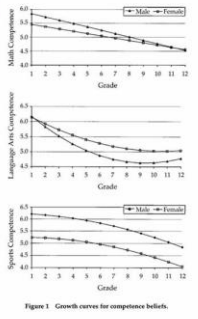


Figure 1 Growth curves for competence beliefs.

Figure 1. Competency beliefs for males and females in math, language arts and sports from grades one through twelve. From "Changes in children's self-competence and values: Gender and domain differences across grades one through twelve," by J.E. Jacobs, S. Lanza, D.W. Osgood, J.S. Eccles and A. Wigfield, 2002, *Child Development, 73* , p.516. Copyright 2010 by Blackwell Publishing. Reprinted with permission.

What can teachers do?

Instruction + | Effort= | Achievement

Students will put forth more effort in class and on tests if:

- Assignments are just challenging enough
- Classroom tests are based on what was taught
- Effort will improve their performance
- Feedback (written and verbal) is directed toward improvement rather than judgment

What can students do?

Instruction + | Effort= | Achievement

Students effort in class is defined by teachers as:

- Paying attention
- Responding to questions
- Completing assignments carefully
- Studying*
- Seek help or ask questions when they don't understand*
- Complete homework*

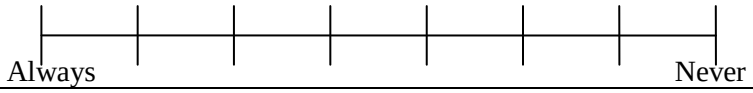
*Increase in importance during middle school and high school

How can students improve effort on PSSA and 4Sight?

- Recognize the importance and value of assessments
- Understand that test scores impact self-efficacy
- Work toward long-term goals
- Improve achievement
- Practice "mindfulness"

Appendix D

Field Notes: Data Coaching Form

❑ Consistently High	❑ Consistently Low	❑ Erratic
Feelings		
What influenced this pattern (class, teacher, personal, friends, subject preference)		
Test taking effort		
What changes, if any, would you like to make to this profile?		
Goal Setting		
Comments		

Student Name

PSSA Data from last year

Reading		Math		Writing		Science	
Score	Level A, P, B, BB	Score	Level A, P, B, BB	Score	Level A, P, B, BB	Score	Level A, P, B, BB

4 Sight Data Summary for this year

	Reading			Math		
	Score	Level	Goal	Score	Level	Goal
Baseline						
Test 1						
Test 2						
Test 3						
Test 4						

Test	How I did in Reading	How I did in Math
Baseline		
1		
2		
3		
4		

Appendix E

Qualitative Data Codes

I/ Importance			
	P /Positive		
		IPP	Parents
		IPT	Teachers/school staff
		IPF	Feedback
		IPC	Consequence
		IPS	PSSA Prep
		IPL	Integral to learning
		IPG	General statements
	N/Negative		
		INT	Teachers/school staff
		INF	Lack of feedback
		INC	Lack of consequence
		ING	General statements

E/Effort			
	P /Positive		
		EPP	Parents
		EPM	Classmates
		EPS	Internal motivation
		EPT	Teacher preparation
		EPB	Ability beliefs
	O/Neutral		
	N/Negative	ENM	Classmates
		ENS	Off task/lack of focus
		ENT	Lack of teacher preparation
		ENB	Low ability beliefs
		ENC	Cost in time or energy
		ENA	Negative attitude
		ENQ	Lack of consequence
		ENG	General statements

F/Feelings			
	P /Positive		
		FPP	Pride
		FPS	Smart
		FPI	Improving
		FPC	Successful/confident
		FPER	Physical response, positive*
	O/Neutral		
		FOC	Wondering/curious
		FOA	Acceptance/resignation
		FOB	Balanced/objective
	N/Negative		
		FND	Self doubt/failure
		FNW	Worry/nervous/scared
		FNP	Disappointed
		FNR	Regret
		FNE	Embarrassment
		FNO	Overwhelmed
		FNA	Anger
		FNF	Frustration
		FNG	General/”bad”
		FNC	Shocked/Negatively surprised
		FNER	Physical response, negative*

*Physical responses were coded when they were noteworthy or outside the norm.